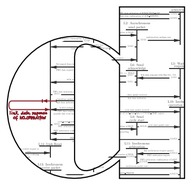
Ethics for Nerds

An Advanced Course in Computer Science
Summer Semester 2020

Current Topics C6.1
Algorithmic Decision-Making
& Algorithmically Supported Decision-Making

Overview



Prof. Holger Hermanns,
Kevin Baum, Sarah Sterz

UNIVERSITÄT
DES
SAARLANDES

https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

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

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

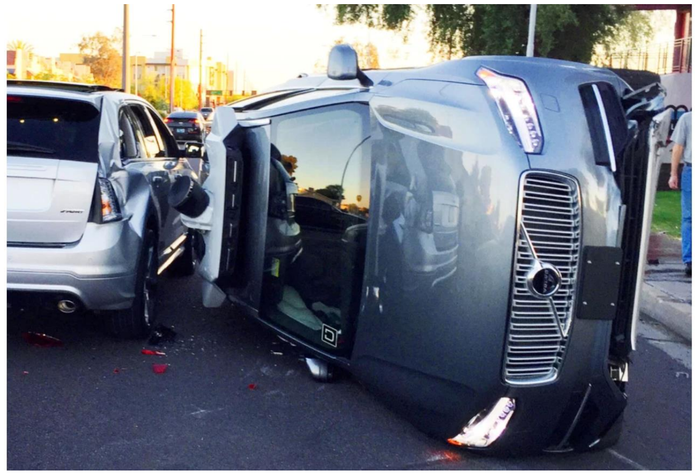
May 23, 2016

https://www.technologyreview.com/2019/05/28/65748/ai-algorithms-liability-human-blame/

MIT Technology Review

Sign in | Subscribe

Topics | Magazine | Newsletters | Events



An image showing the aftermath of a self-driving car accident, with an uber vehicle on its side

TEMPE POLICE DEPARTMENT

Tech policy / AI Ethics

When algorithms mess up, the nearest human gets the blame

A look at historical case studies shows us how we handle the liability of automated systems.

by **Karen Hao**

May 28, 2019

https://www.weforum.org/agenda/2020/02/where-is-artificial-intelligence-going/

WORLD ECONOMIC FORUM Agenda Platforms Reports Events About English TopLink

Global Agenda | Artificial Intelligence and Robotics | Emerging Technologies | Fourth Industrial Revolution

Here are 3 big concerns surrounding AI - and how to deal with them

- Artificial intelligence (AI) needs to be democratized to help more people understand it and embrace its potential;
- We need to develop regulations for AI that are agile and adapt to this rapidly progressing technology;
- A focus on “Trustworthy AI” offers a promising model for innovation and the governance of AI.

17 Feb 2020

Bowen Zhou
President, JD Cloud & AI; Chair, JD Technology Committee; Vice-President, JD.COM

https://www.theverge.com/2020/6/24/21301465/ai-machine-learning-racist-crime-prediction-coalition-critical-technology-springer-study

POLICY SCIENCE TECH

AI experts say research into algorithms that claim to predict criminality must end

AI is in danger of revisiting the pseudoscience of physiognomy

By James Vincent | Jun 24, 2020, 6:45am EDT

SHARE

https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

DER TAGESSPIEGEL

EU guidelines 08.04.2019, 15:48 Uhr

Ethics washing made in Europe

On Tuesday, the EU has published ethics guidelines for artificial intelligence. A member of the expert group that drew up the paper says: This is a case of ethical white-washing. VON THOMAS METZINGER

Thomas Metzinger (Archivbild 2011) FOTO: FREDRIK VON ERICHSEN DPA/LRS

Artificial Intelligence cannot be trustworthy

A VAST FIELD


Interpretable ML
Discrimination
Trust
Human in the Loop
Responsible AI
Bias
Autonomy
AI Ethics
Explainability
Regulation
Ethical White-Washing
Self-Fulfilling Prophecies
Transparency
Trustworthy AI
Fair AI
Math-Washing
Accountability
Responsibility
Watch Dogs
Voids, Gaps, and Diffusion of Responsibility

ADM: Algorithmic Decision-Making

https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf

European Parliament

Understanding algorithmic decision-making: Opportunities and challenges

https://algorithmwatch.org/das-adm-manifest-the-adm-manifesto/

ALGORITHM WATCH

Search...

Newsletter · Events · Contact · Press · · Deutsch

ABOUT / RESEARCH / STORIES / BLOG / DONATE

The ADM Manifesto

Algorithmic decision making (ADM) is a fact of life today; it will be a much bigger fact of life tomorrow. It carries enormous dangers; it holds enormous promise. The fact that most ADM procedures are black boxes to the people affected by them is not a law of nature. It must end.

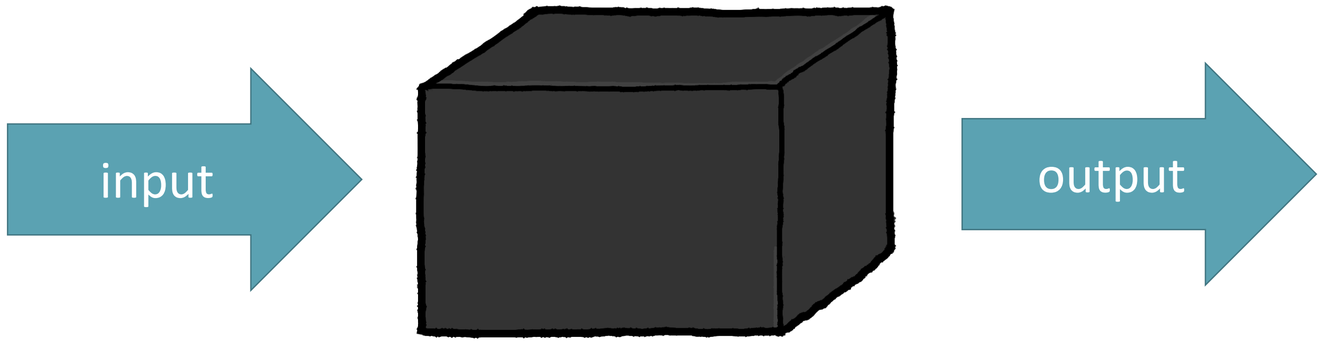
1. ADM is never neutral.
2. The creator of ADM is responsible for its results. ADM is created not only by its designer.
3. ADM has to be intelligible in order to be held accountable to democratic control.
4. Democratic societies have the duty to achieve intelligibility of ADM with a mix of technologies, regulation, and suitable oversight institutions.
5. We have to decide how much of our freedom we allow ADM to preempt.

We call the following process algorithmic decision making (ADM):

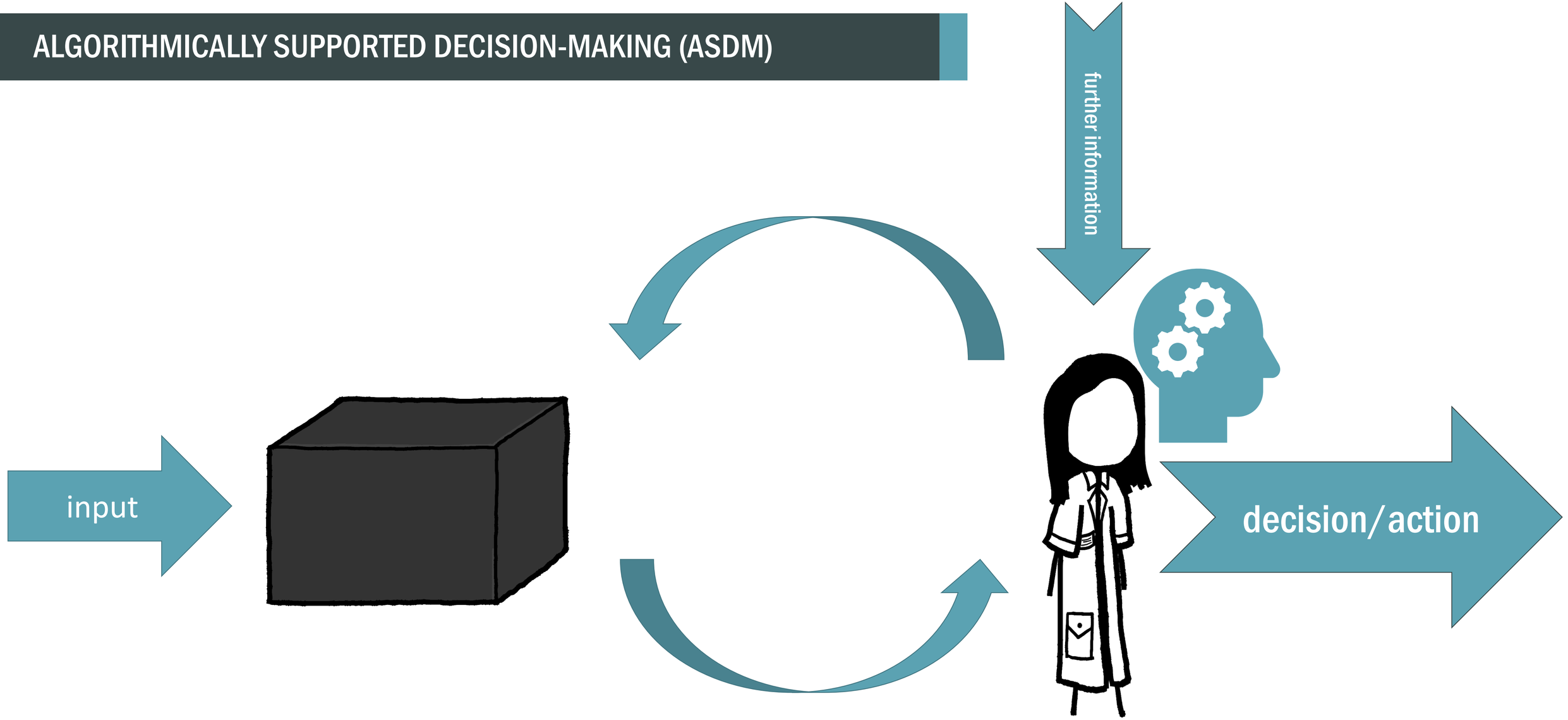
- design procedures to gather data,
- gather data,
- design algorithms to
  - analyse the data,
  - interpret the results of this analysis based on a human-defined interpretation model,
  - and to act automatically based on the interpretation as determined in a human-defined decision making model.

Read our **mission statement**.

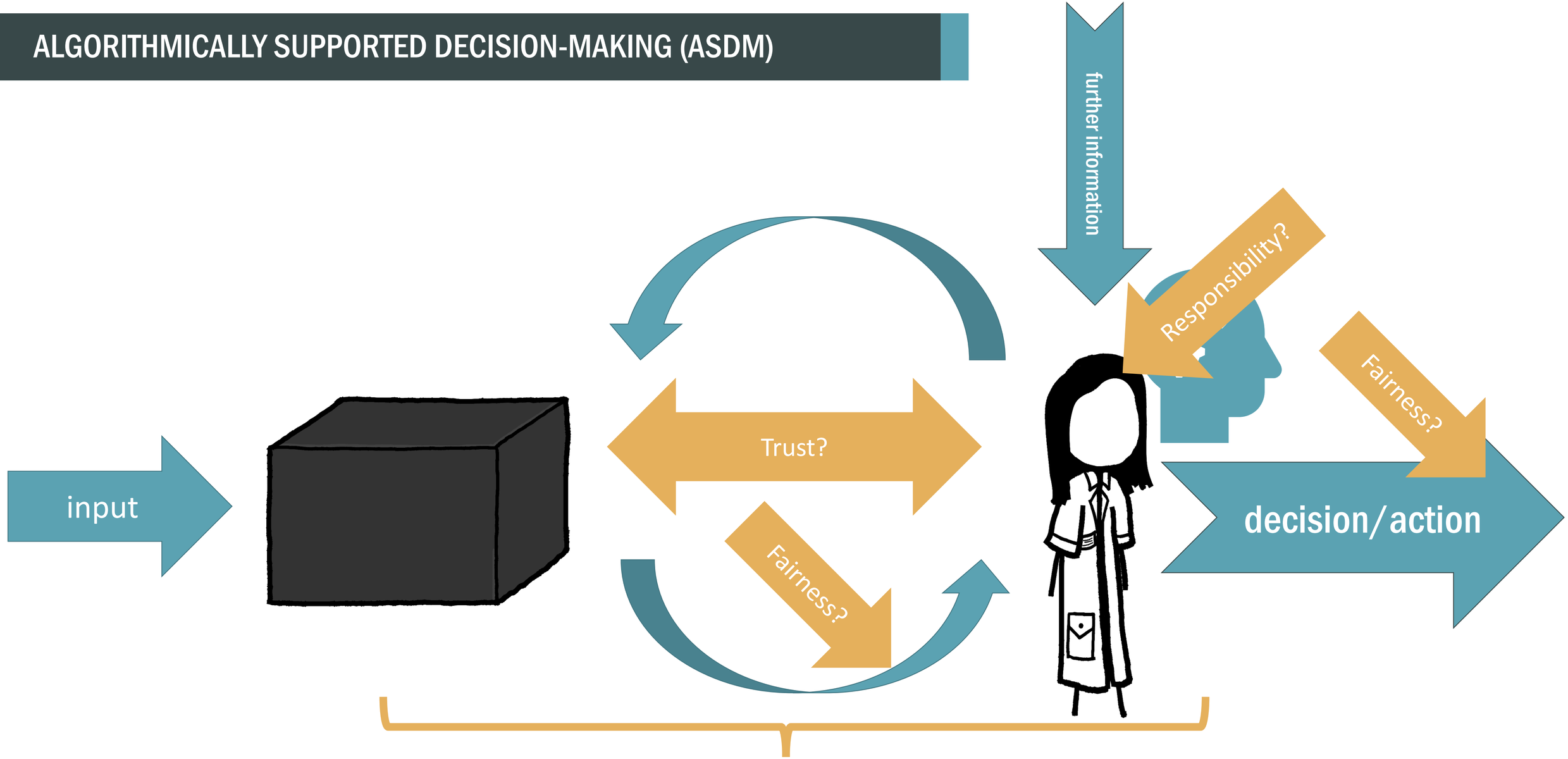
ALGORITHMIC DECISION-MAKING (ADM)



ALGORITHMICALLY SUPPORTED DECISION-MAKING (ASDM)



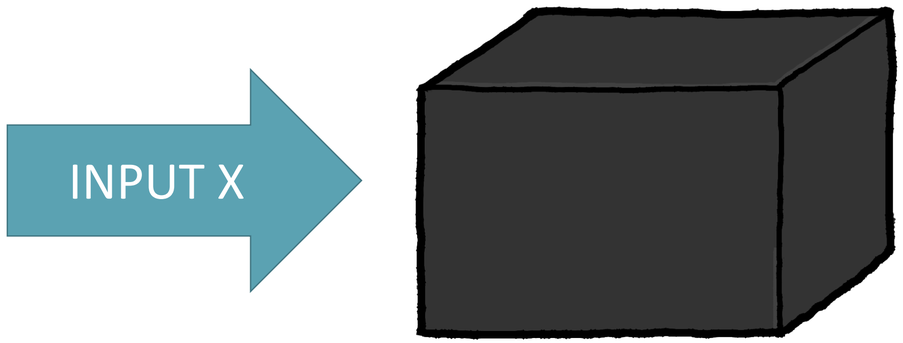
ALGORITHMICALLY SUPPORTED DECISION-MAKING (ASDM)



Explainability, transparency, interpretability, ….

Kinds of Models, the Black Box Perspective, and some Words on Feedback Loops

BLACK BOX PERSPECTIVE & “MODEL AGNOSTICISM”

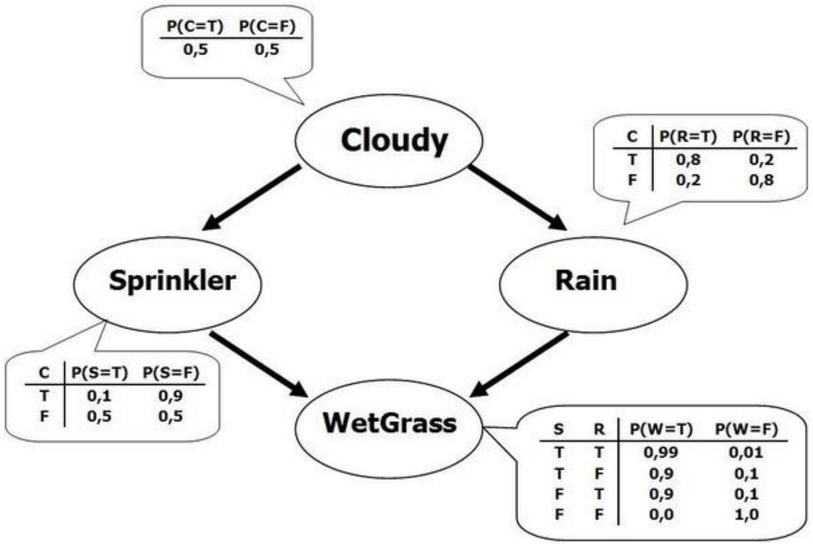


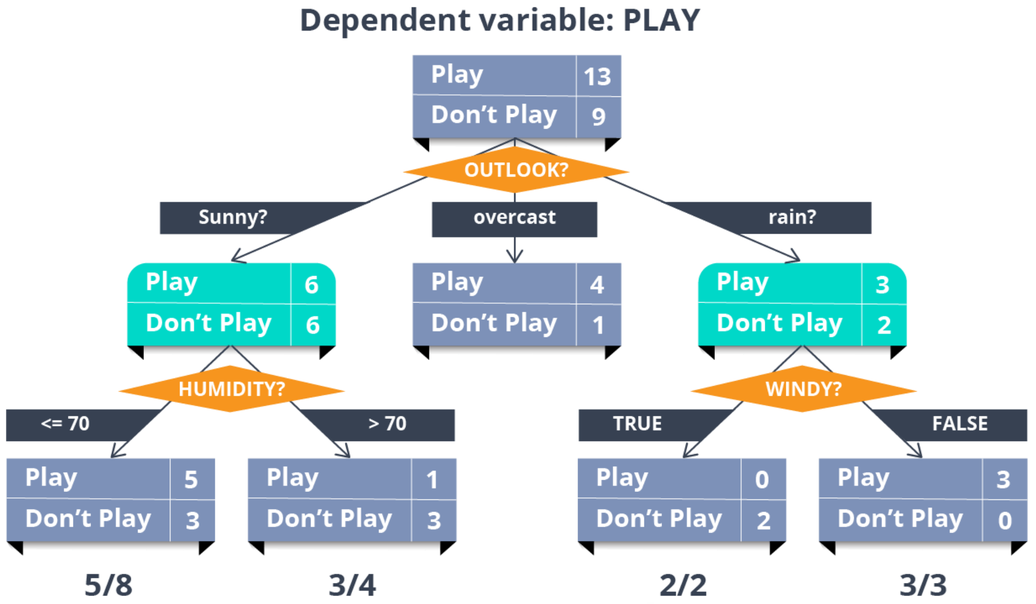
Example: classification. Y=…

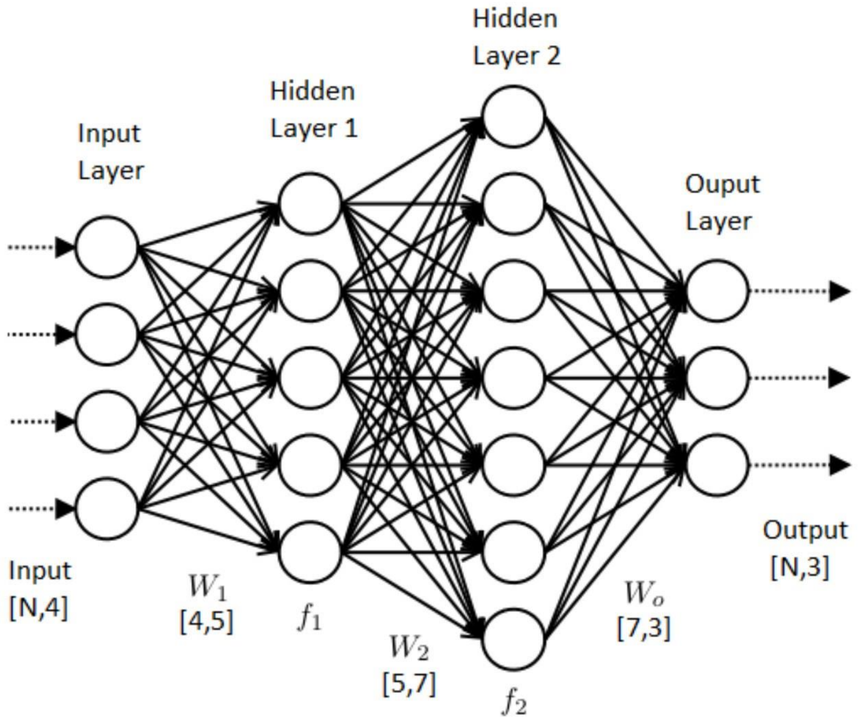
- … X belongs to cat 1
- … X belongs to cat 2
- …

- Sometimes, this perspective is taken for methodological reasons.
- Sometimes, we do not have another perspective at hand (CNNs etc.)
- Keyword: non-/sub-symbolic representations…

```
53
54      uppercase_sample = 'ABCDEFGHIJKLMNOPQRSTUVWXYZ'
55      lowercase_sample = 'abcdefghijklmnopqrstuvwxyz'
56      digit_sample = '0123456789'
57      if keras.backend.image_dim_ordering() != 'tf' and count <= 100:
58          keras.backend.set_image_dim_ordering('tf')
59          print("INFO: '~/.keras/keras.json' sets 'image_dim_ordering' to "
60                "'th', temporarily setting to 'tf'")
61
62      # Create TF session and set as Keras backend session
63      sess = tf.Session()
64      keras.backend.set_session(sess)
65
66      # Get MNIST test data
67      X_train, Y_train, X_test, Y_test = data_mnist(train_start=train_start,
68                                                    train_end=train_end,
69                                                    test_start=test_start,
70                                                    test_end=test_end)
71
72      assert Y_train.shape[1] == 10
```

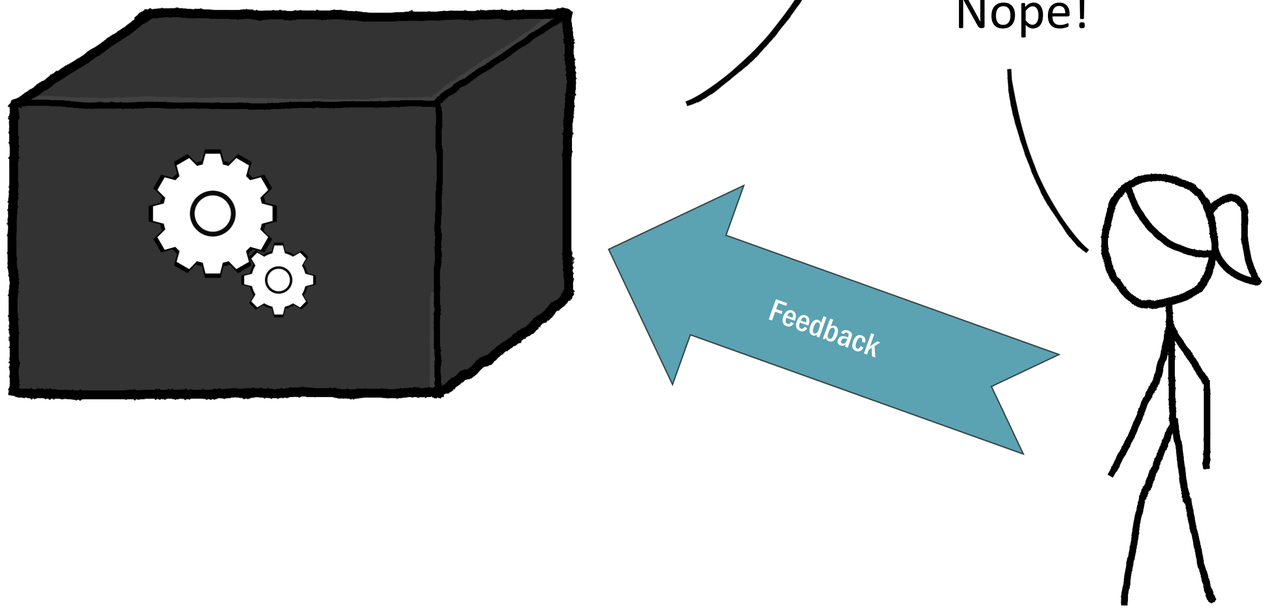






https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

THE FEEDBACK LOOP YOU ALL KNOW

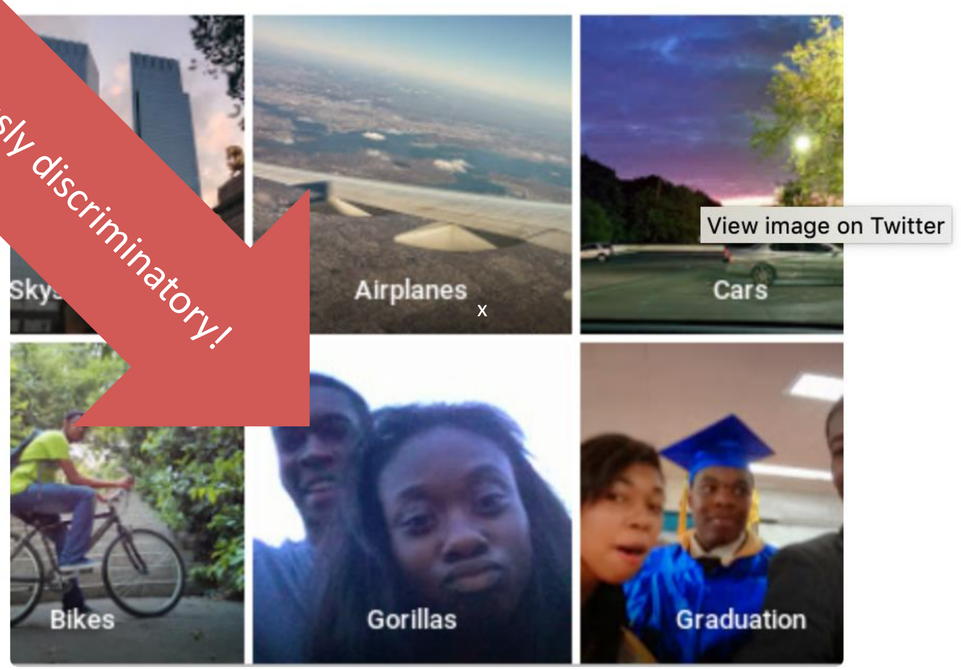


seriously discriminatory!

https://twitter.com/jackyalcine/status/615329515909156865

jackyalcine SOON TwitchCon 2019
@jackyalcine

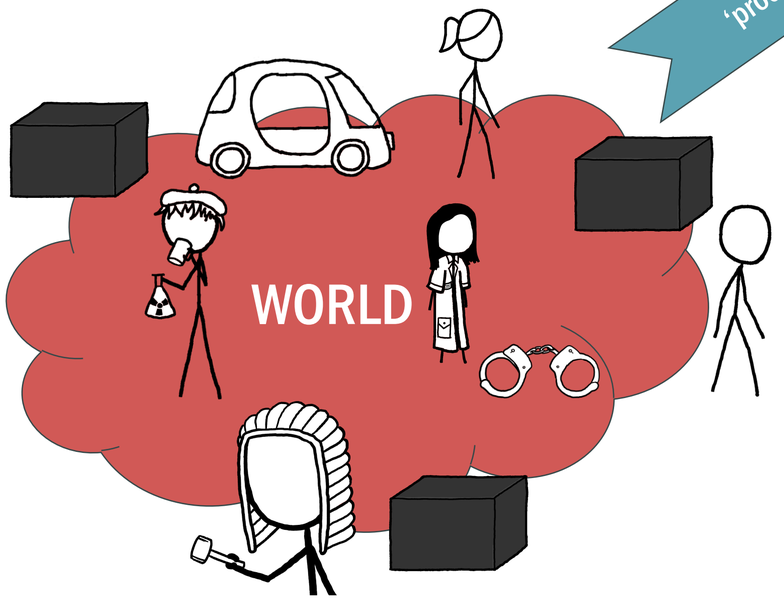
Google Photos, y'all fucked up. My friend's not a gorilla.



2,757 3:22 AM - Jun 29, 2015

3,663 people are talking about this

THE NOT SO OBVIOUS ”FEEDBACK LOOP”



‘produces’

“BIG DATA”

grist for

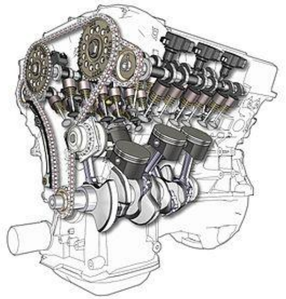
machine learning algorithms

results in

model

becomes efficacious causal part of

A MISLEADING ANALOGY: “Data is the new oil”

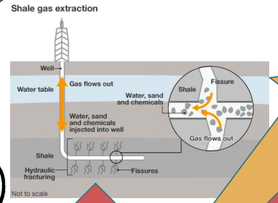


“BIG DATA”

grist for

machine learning algorithms

‘produces’



results in

becomes efficacious causal part of

model

FOCUS

General Problems

Embedded Values,
Math-Washing,
Pseudo-Solutions,
Self-Fulfilling Prophecies

Discrimination
& Algorithmic Fairness

Responsibility
& Explainability

(significantly shortened due to the current crisis)

Ethics for Nerds

An Advanced Course in Computer Science
Summer Semester 2020

Current Topics C6.2
Algorithmic Decision-Making
& Algorithmically Supported Decision-Making

Examples, Embedded Values, the Ground Truth Problem

Prof. Holger Hermanns,
Kevin Baum, Sarah Sterz

UNIVERSITÄT
DES
SAARLANDES

FOCUS

General Problems

Embedded Values,
Math-Washing,
Pseudo-Solutions,
Self-Fulfilling Prophecies

Discrimination
& Algorithmic Fairness

Responsibility
& Explainability

(significantly shortened due to the current crisis)

FOCUS

General Problems

Embedded Values,
Math-Washing,
Pseudo-Solutions,
Self-Fulfilling Prophecies

Discrimination
& Algorithmic Fairness

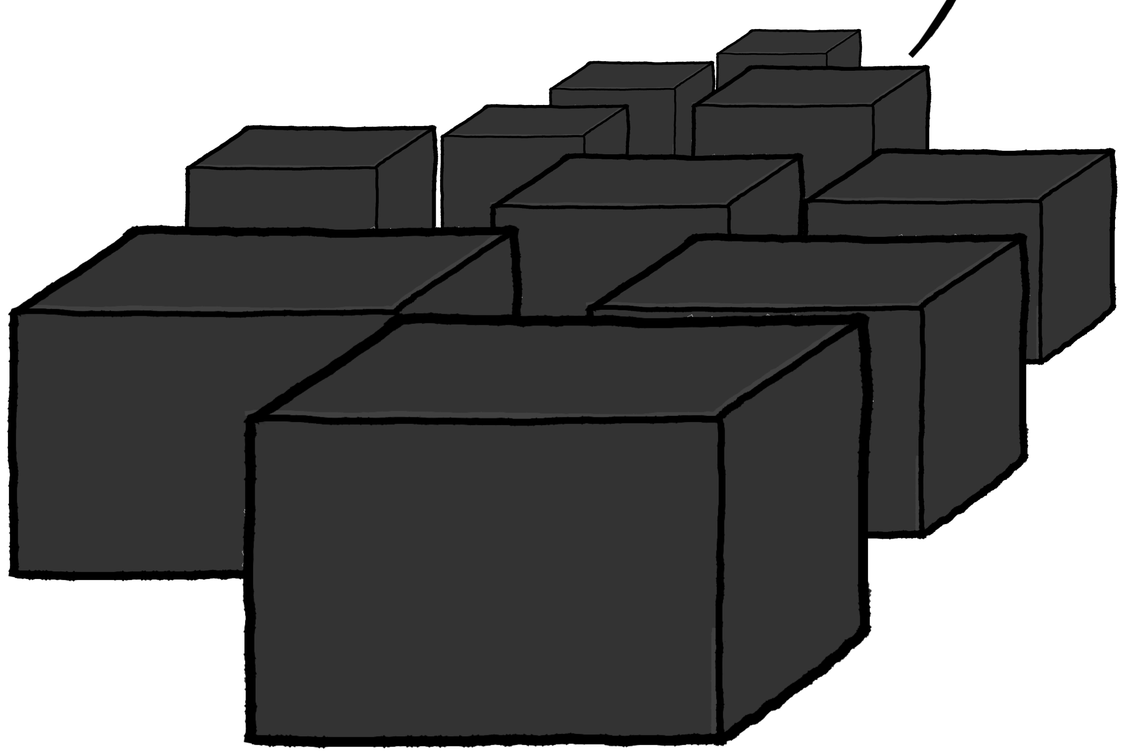
Responsibility
& Explainability

(significantly shortened due to the current crisis)

Some Examples

ML AS A PREDICTOR AND DECIDER

- imprisonment/probation decisions
- jobs
- crime prediction
- assessment of creditworthiness
- checking of alimony payment
- blocking of credit cards
- terror prediction
- refusal of visa
- ...



WE ARE BIASED

Lunch Luck

http://www.nature.com/news/2011/110411/full/news.2011.227.html

Overlooked factors in the analysis of parole decisions

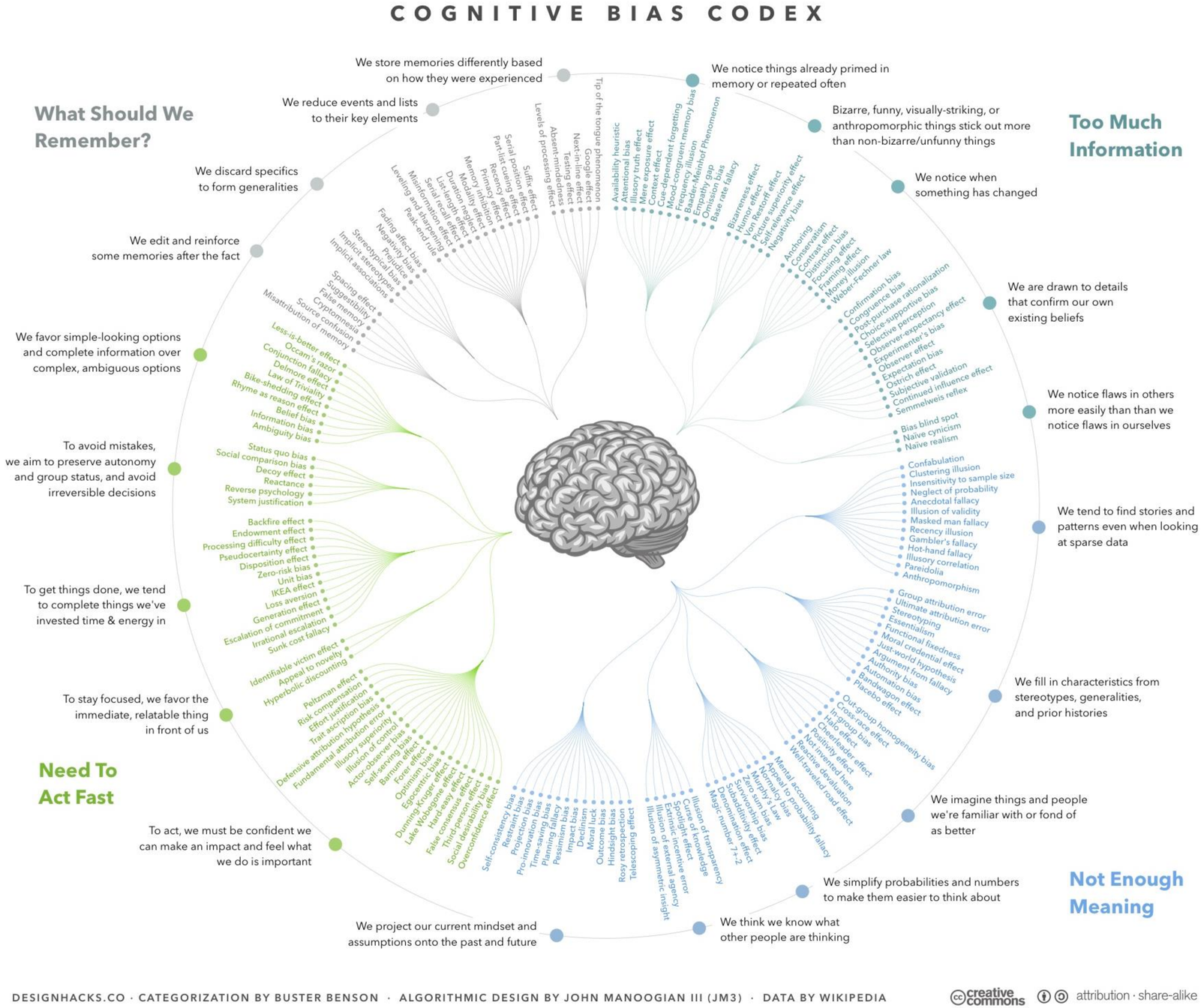
Keren Weinshall-Margel and John Shapard

PNAS October 18, 2011 108 (42) E833; https://doi.org/10.1073/pnas.1110910108

Article | Info & Metrics | PDF

Danziger et al. (**1**) concluded that meal breaks taken by Israeli parole boards influence the boards' decisions. This conclusion depends on the order of cases being random or at least exogenous to the timing of meal breaks. We examined data provided by the authors and obtained additional data from 12 hearing days (n = 227 decisions).* We also interviewed three attorneys, a parole panel judge, and five personnel at Israeli Prison Services and Court Management, learning that case ordering is not random and that several factors contribute to the downward trend in prisoner success between meal breaks. The most important is that the board tries to complete all cases from one prison before it takes a break and to start with another prison after the break. Within each session, unrepresented prisoners usually go last and are less likely to be granted parole than prisoners with attorneys. Using the same decision rules as Danziger et al., our data indicate that unrepresented prisoners account for about one-third of all cases, but they prevail only 15% of the time, whereas prisoners with counsel prevail at a 35% rate.

This nonrandom order of cases might have become apparent had the authors not limited their analysis. They lumped together decisions rejecting parole and cases that were deferred to a later date. Theoretically and in practice, deferrals are not comparable to rejections of parole.



https://en.wikipedia.org/wiki/List_of_cognitive_biases
(unsurprisingly, it's quite long...)

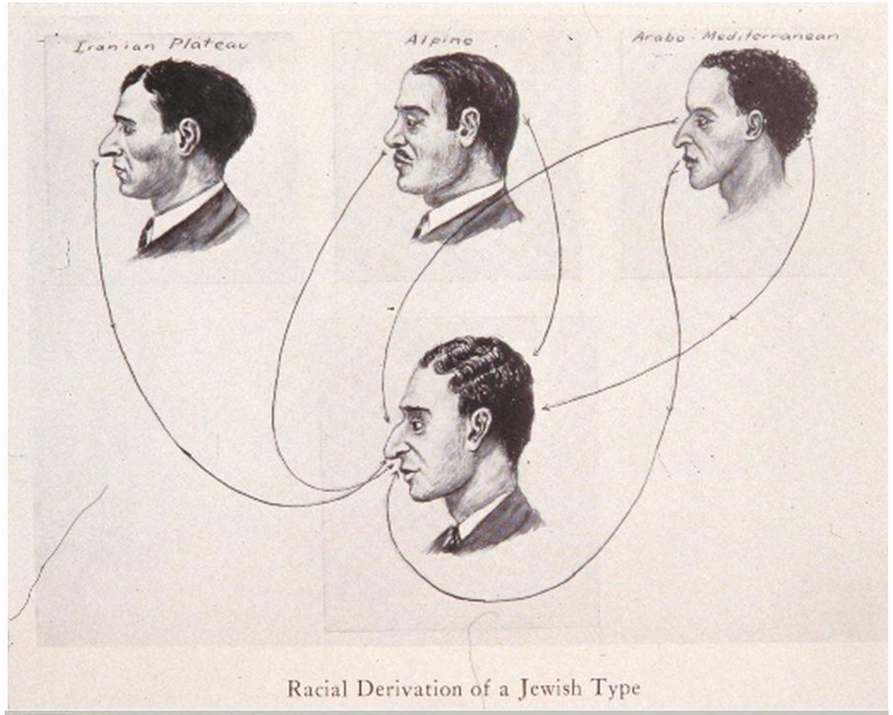
COMPAS AS EXAMPLE

https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/
https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)

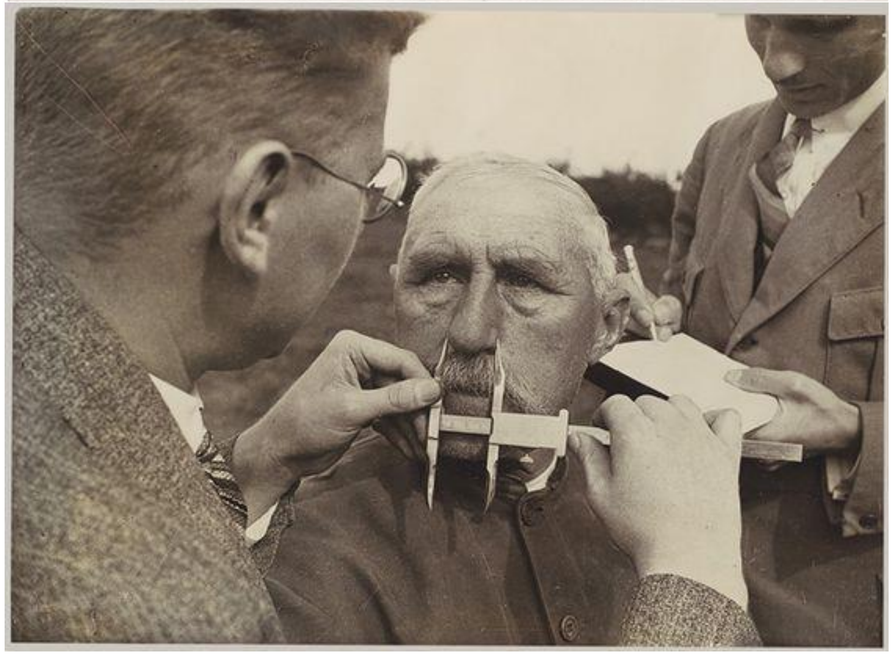
- proprietary, closed-source risk assessment algorithm used in court
- Eric L. Loomis was sentenced to six years of prison, because the algorithm said he had “a high risk of violence, high risk of recidivism, high pretrial risk.”
- Loomis challenged the sentence, but was rejected:
  - “The key to our product is the algorithms, and they’re proprietary[…] We’ve created them, and we don’t release them because it’s certainly a core piece of our business.”
  - The court reasoned that knowledge of the algorithm's output was a sufficient level of transparency
  - Loomis knew everything the court knew – as judges do not know the internals of the algorithm, either.
  - And: The judges were sure that, in this case, the sentence would’ve been the same with or without COMPAS’ assessment.

https://www.wsj.com/articles/wisconsin-supreme-court-to-rule-on-predictive-algorithms-used-in-sentencing-1465119008



Racial Derivation of a Jewish Type

From Earnest Hooton, "The Twilight of Man", 1939, http://i.imgur.com/X2lHfiD.jpg

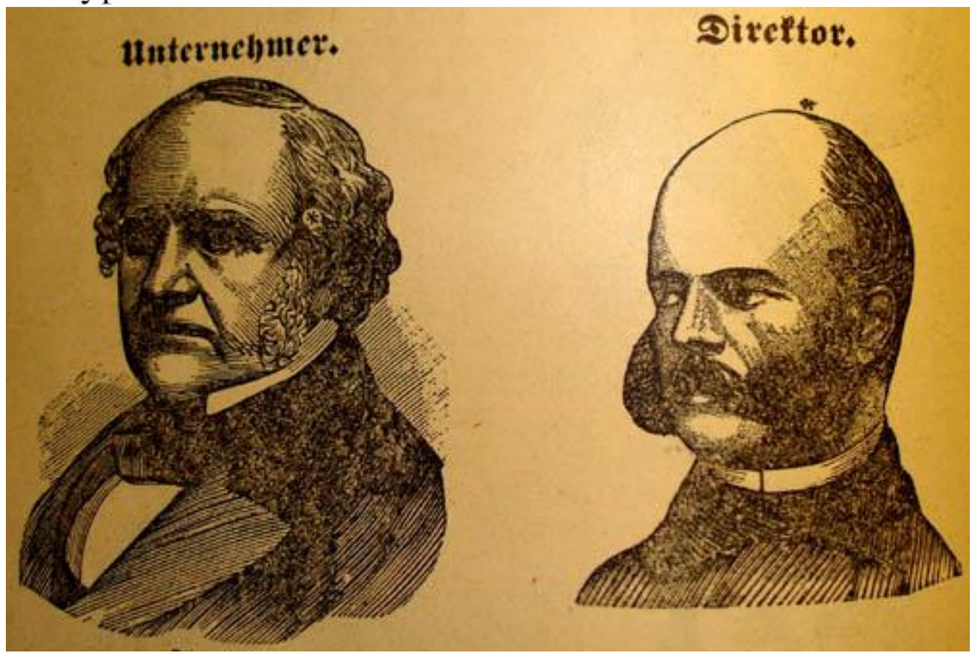


https://de.pinterest.com/pin/141230138285293618/

(a)

(b)

Figure 9. (a) The four subtypes of criminal faces; (b) The three subtypes of non-criminal faces.



https://upload.wikimedia.org/wikipedia/de/4/42/Phrenologie3.jpg

“GAYDAR"

https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph

Artificial intelligence (AI)

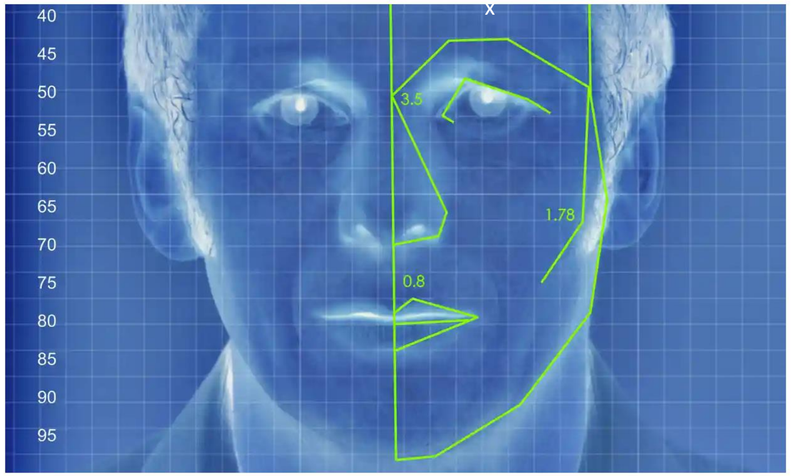
This article is more than **2 years old**

New AI can guess whether you're gay or straight from a photograph

An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions

Sam Levin *in San Francisco*
@SamTLevin Email

Fri 8 Sep 2017 00.46 BST



▲ An illustrated depiction of facial analysis technology similar to that used in the experiment. Illustration: Alamy

The study: https://www.gsb.stanford.edu/sites/gsb/files/publication-pdf/wang_kosinski.pdf

https://en.wikipedia.org/wiki/Capital_punishment_for_homosexuality

Capital punishment for homosexuality

From Wikipedia, the free encyclopedia

Capital punishment for homosexuality was historically implemented by a numbe
Being prescribed by the law does not necessarily mean that the penalty is carried o

Contents [hide]

In current state laws [edit]

Further information: LGBT rights by country or territory, LGBT in Islam, and Sha

As of 2020, the following jurisdictions prescribe the death penalty for homosexuality

- Afghanistan. A new Penal Code enacted in February 2018 explicitly crimina appropriate punishment if homosexua‘ acts could be proven".[1] The sharia cate punishment of stoning, when strict evidential requirements are met. The Hanafi penalty for a number of reasons. No known death sentences for homosexuality
- Brunei's *Sharia Penal Code*, implemented in stages since 2014, prescribes facto" moratorium on the execution of the death penalty has been in force in the
- Iran.[6] Homosexual intercourse is declared a capital offense in Iran's *Islam* homosexual behaviour in 2005-2006 and in 2016, in some cases on dubious ch
- Mauritania.[6] According to a 1984 law, Muslim men can be stoned for enga since 1987.[10]
- Nigeria, where several northern states have adopted sharia-based criminal
- Pakistan, where the death penalty for homosexual acts is technically permi
- Qatar, applicable only to Muslims, for extramarital sex regardless of the ger adults and in private.[12]
- Saudi Arabia, which does not have codified criminal laws.[6] According to th death.[9] There were unconfirmed reports that two cross-dressing Pakistani nati
- Somalia (Jubaland), (Somaliland) where Islamic courts have impo
- Sudan, for a third conviction.[6]
- United Arab Emirates: Legal experts disagree on whether the federal law o whoever used coercion in having sexual intercourse with a female or sodomy wi the UAE. Although application of Zina provisions in the UAE appears to be limite penalty.[14] A recent Amnesty International report states that they are not aware
- Yemen: If the same-sex activity occurs outside of marriage, the death pena

“GAYDAR” DEBUNKED

https://medium.com/@blaisea/do-algorithms-reveal-sexual- orientation-or-just-expose-our-stereotypes-d998fafdf477

Medium | Equality

Do algorithms reveal sexual orientation or just expose our stereotypes?

Blaise Aguera y Arcas Follow
Jan 11, 2018 · 15 min read

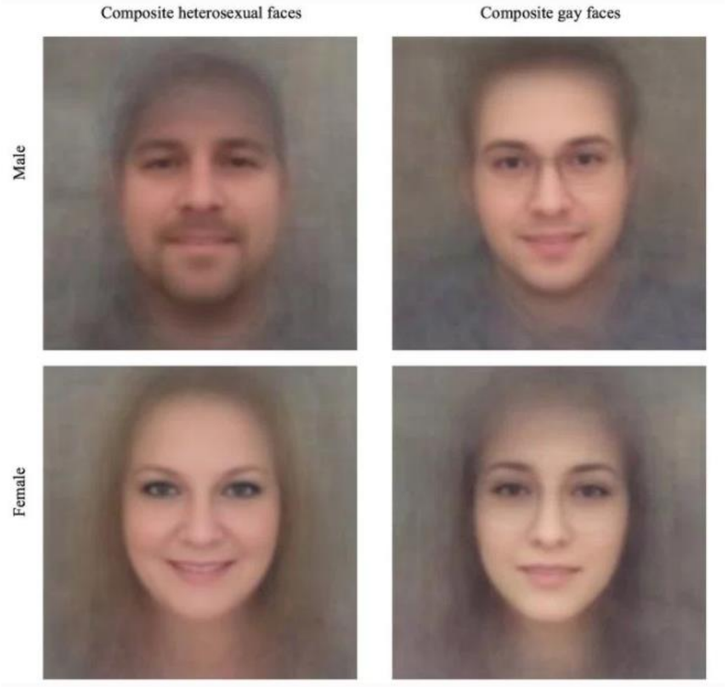
by Blaise Agüera y Arcas, Alexander Todorov and Margaret Mitchell

A study claiming that artificial intelligence can infer sexual orientation from facial images caused a media uproar in the Fall of 2017. The Economist featured this work on the cover of their September 9th magazine; on the other hand two major LGBTQ organizations, The Human Rights Campaign and GLAAD, immediately labeled it “junk science”. Michal Kosinski, who co-authored the study with fellow researcher Yilun Wang, initially expressed surprise, calling the critiques “knee-jerk” reactions. However, he then proceeded to make even bolder claims: that such AI algorithms will soon be able to measure the intelligence, political orientation, and criminal inclinations of people from their facial images alone.

In summary, we have shown how the obvious differences between lesbian or gay and straight faces in selfies relate to grooming, presentation, and lifestyle — that is, differences in culture, not in facial structure. These differences include:

- Makeup
- Eyeshadow
- Facial hair
- Glasses
- Selfie angle
- Amount of sun exposure.

We’ve demonstrated that just a handful of yes/no questions about these variables can do nearly as good a job at guessing orientation as supposedly sophisticated facial recognition AI. Further, the current generation of facial recognition remains sensitive to head pose and facial expression. Therefore — at least at this point — it’s hard to credit the notion that this AI is in some way superhuman at “outing” us based on subtle but unalterable details of our facial structure.



“GAYDAR"

https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph

Support The Guardian
Available for everyone, funded by readers
Contribute → Subscribe →

Search jobs | Sign in | Search | The Guardian | International edition

News | Opinion | Sport | Culture | Life

World UK Science Cities Global development Football **Tech** Business Environm

Artificial intelligence (AI)

This article is more than **2 years old**

New AI can g… straight f…

Sam Levin *in San Francisco*
@SamTLevin Email
Fri 8 Sep 2017 00.46 BST

▲ An illustrated depiction of facial analysis technology similar to that used in the experiment. Illustration: Alamy

The study: https://www.gsb.stanford.edu/sites/gsb/files/publication-pdf/wang_kosinski.pdf

Does it make the model better or less worrisome?

(No, but it becomes more about mass surveillance and less about AI)

ISSUE 1: BUILT-IN VALUES THROUGH INTERPRETATION

Beware to implicitly (and maybe inadvertently) implement (social or moral) values into your system!

COMPAS AS EXAMPLE FOR EMBEDDED VALUES

https://assets.documentcloud.org/documents/2702103/Sample-Risk-Assessment-COMPAS-CORE.pdf

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)

32. If you lived with both parents and they later separated, how old were you at the time?
☑ Less than 5 ☐ 5 to 10 ☐ 11 to 14 ☐ 15 or older ☐ Does Not Apply

33. Was your father (or father figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

34. Was your mother (or mother figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

Please think of your friends and the people you hung out with in the past few (3-6) months.

39. How many of your friends/acquaintances have ever been arrested?
☐ None ☐ Few ☑ Half ☐ Most

64. Do you have an alias (do you sometimes call yourself by another name)?
☑ No ☐ Yes

COMPAS AS EXAMPLE FOR EMBEDDED VALUES

https://assets.documentcloud.org/documents/2702103/Sample-Risk-Assessment-COMPAS-CORE.pdf

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)

78. How strongly do you agree or disagree with the following: I always behaved myself in school?
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

95. How often did you feel bored?
☐ Never ☑ Several times/mo ☐ Several times/wk ☐ Daily

Criminal Personality

The next few statements are about what you are like as a person, what your thoughts are, and how other people see you. There are no 'right or wrong' answers. Just indicate how much you agree or disagree with each statement.

112. "I am seen by others as cold and unfeeling."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

137. "Some people just don't deserve any respect and should be treated like animals."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

ISSUE 2: BLIND HUNTING IN BAD DATA SETS

Just because there is a labeled data set
it doesn’t mean that it is an appropriate data set.

Always think about the source of the data and whether what you do makes sense at all.

A GOOD EXAMPLE FOR A BAD PRACTICE

https://arxiv.org/abs/1611.04135

***Automated Inference on Criminality using Face Images* by Wu and Zhang**

- Evaluates an AI approach of
  - phrenology = link personality and character to head shape
  - physiognomy = link personality and character to facial features

Very popular in the 19th century; both are regarded as unscientific and are (historically) related to racism; AI/ML has brought a revival of interest in these fields.

- One of many examples, where the paper gives hints for its own narrowness:
  - “The two manifolds consisting of criminal and non-criminal faces appear to be concentric, with the non-criminal manifold lying in the kernel with a smaller span, exhibiting a law of normality for faces of non-criminals. In other words, the faces of general law-biding public have a greater degree of resemblance compared with the faces of criminals, or criminals have a higher degree of dissimilarity in facial appearance than normal people.”

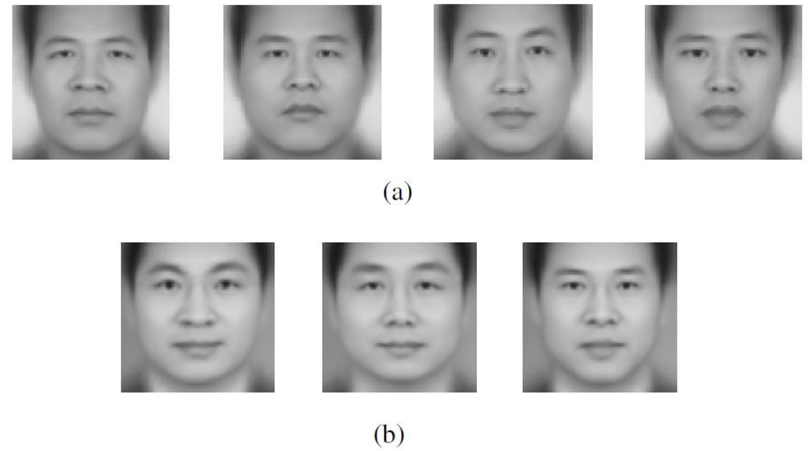


Figure 9. (a) The four subtypes of criminal faces; (b) The three subtypes of non-criminal faces.

CALL FOR MORATORIUM AND BAN

https://www.theverge.com/2020/6/24/21301465/ai-machine-learning-racist-crime-prediction-coalition-critical-technology-springer-study

POLICY SCIENCE TECH

AI experts say research into algorithms that claim to predict criminality must end

AI is in danger of revisiting the pseudoscience of physiognomy

By James Vincent | Jun 24, 2020, 6:45am EDT

SHARE

IS THERE A GROUND-TRUTH? CAN WE ACCESS IT, IF IT EXISTS?

- What is the objective/inter-subjective definition of “being a criminal”?
- What is the objective/inter-subjective definition of “being a good/valuable employee”?
- What is the objective/inter-subjective definition of “being gay/cis”?
- What is the objective/inter-subjective definition of “having a high risk of being a reoffender”?
- …

If there are no such objective/inter-subjective definitions, what does this entail?

If some exist, can we reliably (epistemically) access whether the conditions are given in a case? If not, what does it entail?

(For us computer scientists, but also for our general everyday practices?)

The Idea of Mathwashing

Mathwashing can be thought of using math terms (algorithm, model, etc.) to paper over a more subjective reality. [...]
So I coined "mathwashing" in an attempt to describe the tendency by technologists (and reporters!) to use the objective connotations of math terms to describe products and features that are probably more subjective than their users might think.

Fred Benenson

http://technical.ly/brooklyn/2016/06/08/fred-benenson-mathwashing-facebook-data-worship/

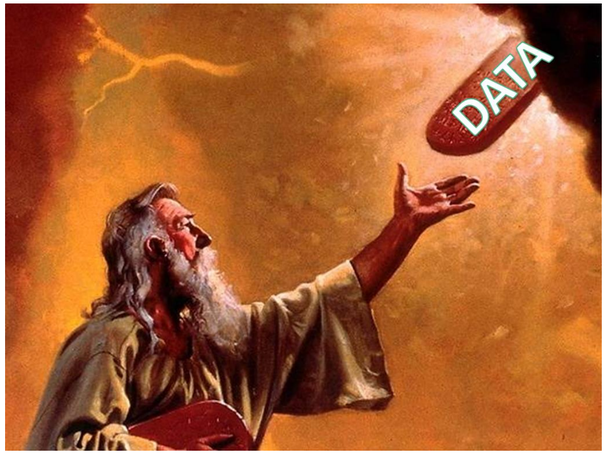
MATHWASHING

What's Wrong with Mathwashing

There's no such thing as perfect data, and if you collect data incorrectly, no math can prevent you from making a bad or dangerous product. In other words, anything we build using data is going to reflect the biases and decisions we make when collecting that data. So I think if we want to "stick to the numbers", we have to be intellectually honest about understanding how we recorded those numbers. Who recorded them and what criteria did they use?

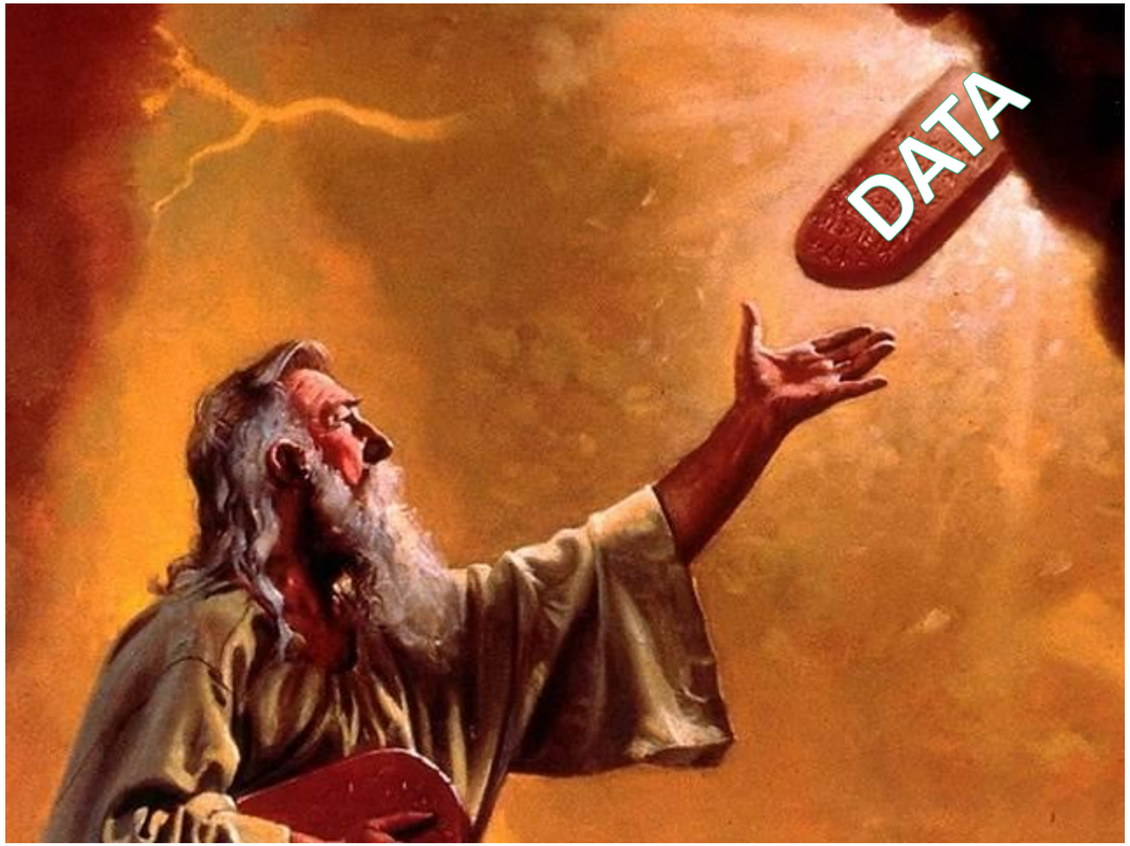
Fred Benenson

http://technical.ly/brooklyn/2016/06/08/fred-benenson-mathwashing-facebook-data-worship/



THE GROUND TRUTH PROBLEM

- Developers need *some data as ground truth*.
- But data for real world problems is *not* given by an unbiased, objective entity!
- Data for real world problems is always *generated* by the real world, more specifically by their inhabitants.
- And so the real world *leaves its marks in the data*.
- (Worse: The relevant data often is not available. But *proxies* are: When we cannot measure X which is a good predictor of Y, we may still have access to Z, which correlates with X.
  I.e. Steps per day (Z), healthy living (X), health (Y).
  Or race (Z) and social situation (X) recidivism rate (Y).)



THE GROUND TRUTH PROBLEM: CRIMINALS

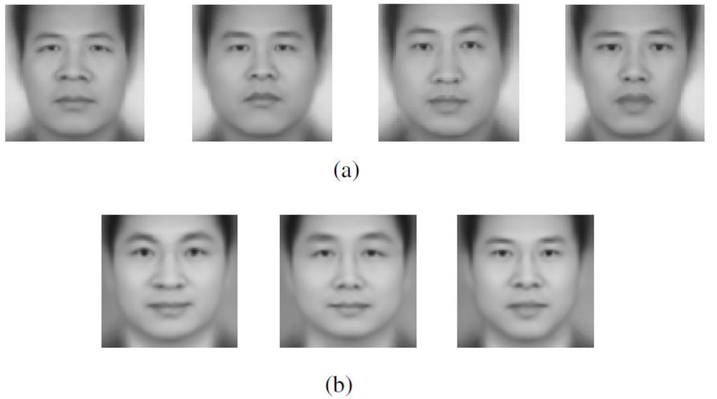


Figure 9. (a) The four subtypes of criminal faces; (b) The three subtypes of non-criminal faces.

Assume you have a list of “Criminals”.
Under which condition is someone a criminal? Sources of trouble:

- What is a crime is depending on the country your data is coming from.
- Who makes it to the list does not only depend on whether someone commits a crime, but
  - … whether he or she was prosecuted.
  - … whether he or she was caught.
  - … whether he or she was convicted.

  Many factors can play a causal role here:
  race, socio-economic status, address, the type of crime,
  even your specific physiognomy, special features….

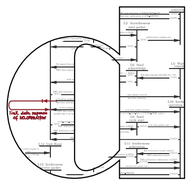
In other words: Whatever the faces above are subtypes of… we do not really know!

Ethics for Nerds

An Advanced Course in Computer Science
Summer Semester 2020

Current Topics C6.3
Algorithmic Decision-Making
& Algorithmically Supported Human Decision-Making

The Contrast to the “Good Old Days”™



Prof. Holger Hermanns,
Kevin Baum, Sarah Sterz

UNIVERSITÄT
DES
SAARLANDES

What makes ADM/ASDM more problematic than the existing, non-algorithmic practices?

Slaves of Statistic Significance

- Intuitively, you would expect that the sanctions for Eric Loomis crimes (directly) depend upon his *own* behavior, attitude and choices.
- If a verdict is solely based on COMPAS, the verdict is solely based on the behavior, attitude and choices of other (former) offenders and how Loomis relates to that.
- Missing qualitative information? What is the role of empathy and intuition?
- Maybe good reasons to keep the human in the loop (but may entail the necessity of a certain degree of understanding of the algorithm.)

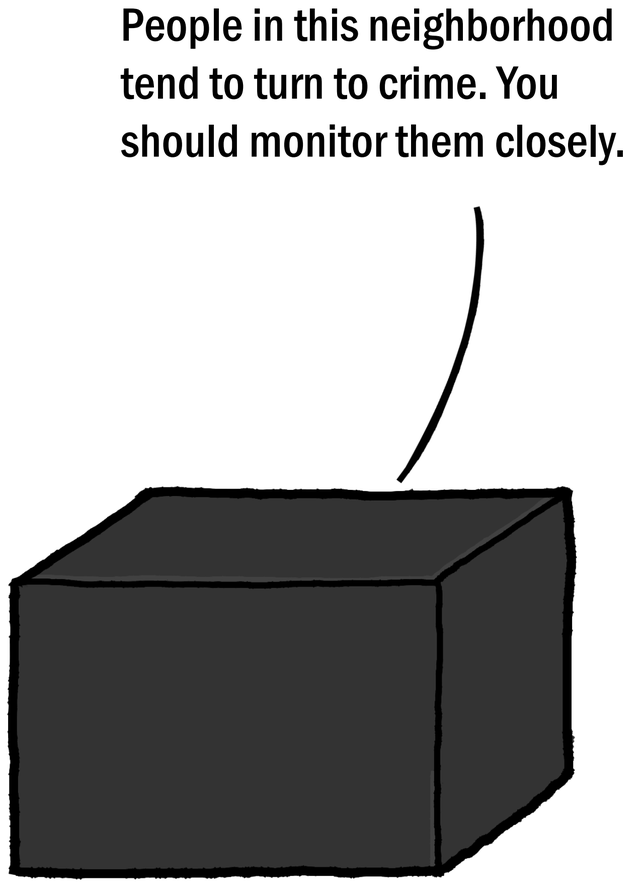
https://www.wsj.com/articles/wisconsin-supreme-court-to-rule-on-predictive-algorithms-used-in-sentencing-1465119008

ANOTHER PROBLEM WITH FEEDBACK LOOPS: SELF-FULFILLING PROPHECIES

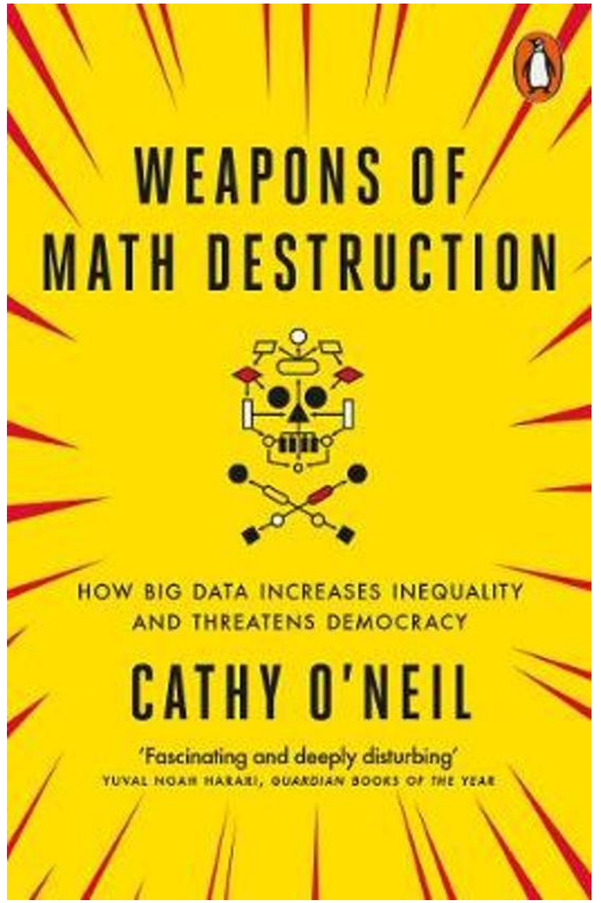
Self-fulfilling Prophecies

- "The self-fulfilling prophecy is, in the beginning, a false definition of the situation evoking a new behavior which makes the original false conception come true." (from *The Self-Fulfilling Prophecy* by Robert K. Merton)
- famous example from sociology: stereotype threat and stereotype boost/lift
- applied to ADM/ASDM contexts:
  - a neighborhood, which is monitored more closely, could become more 'criminal' (at least statistically)
    → self enhancement loop
  - someone who does not get a job because of an alleged risk of getting depressed will more likely become depressed (→ discussed later)

Self-fulfilling prophecies can seriously harm + they are rather hard to predict (ex ante epistemic challenge) and identify (ex post epistemic challenge)!



WEAPONS OF MATH-DESTRUCTION



In her 2016 book *Weapons of Math-Destruction*, Cathy O'Neil identifies three characteristics of especially harmful models, so-called "Weapons of Math-Destruction" ("WMDs"):

- **Opacity**: The model is opaque, and it is acting in way invisible to the affected.
  (→ next week a little bit more on this)
- **Scale**: The model does scale (or, more precisely, its application).
- **Damage**: The model does damage to people (or, more precisely, its application).

Additional important feature: The implementation of **the feedback loop**

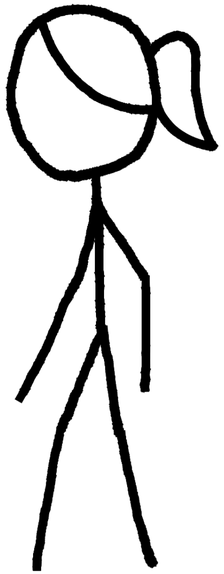
- Are wrong and right predictions evaluated?
- If so, how?
  This is not easy: If I reject an applicant A for some job and accept another candidate B based on a model's recommendation, how do I evaluate that A was not better suited than B?
- Are the evaluations fed back into the system in order to improve its performance?

WE ARE BIASED

Rudolph the Racist

- That's Rudolph the racist, who happens to be a personnel manager.

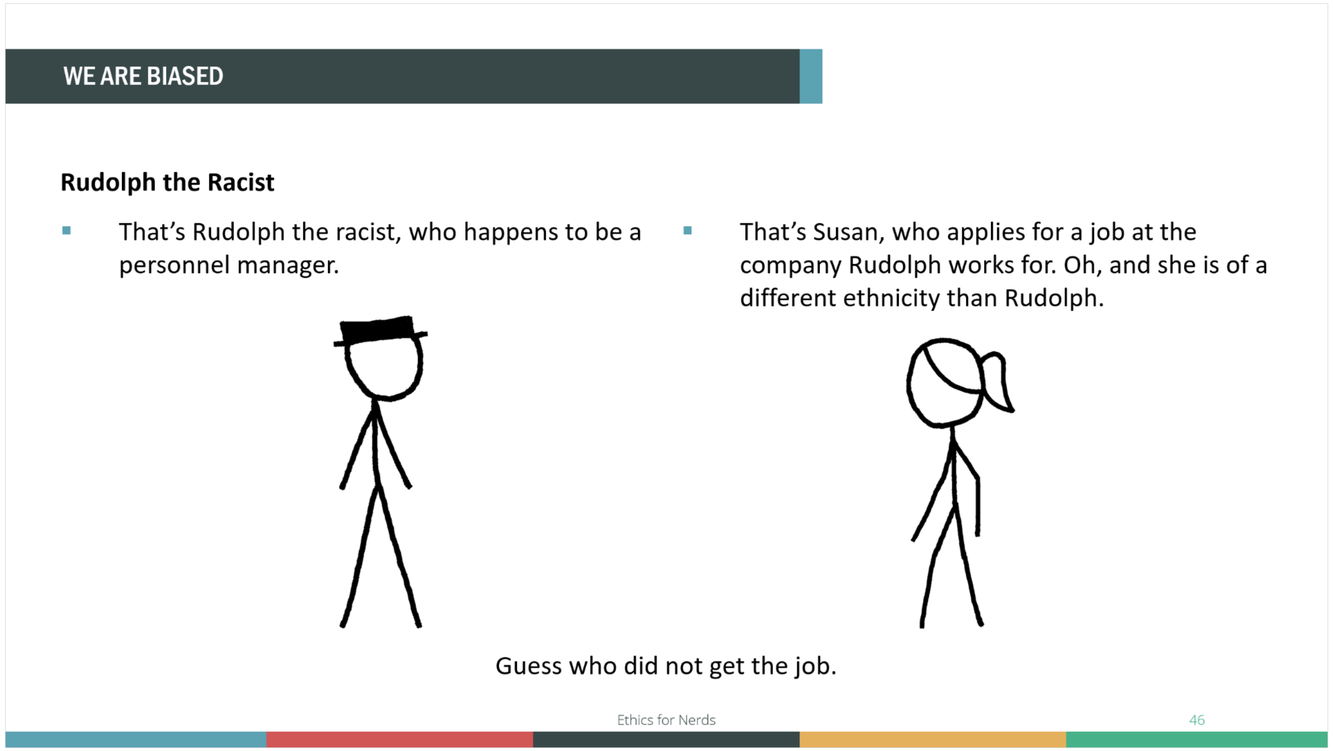
- That's Susan, who applies for a job at the company Rudolph works for. Oh, and she is of a different ethnicity than Rudolph.



Guess who did not get the job.

WEAPONS OF MATH-DESTRUCTION: SCALE

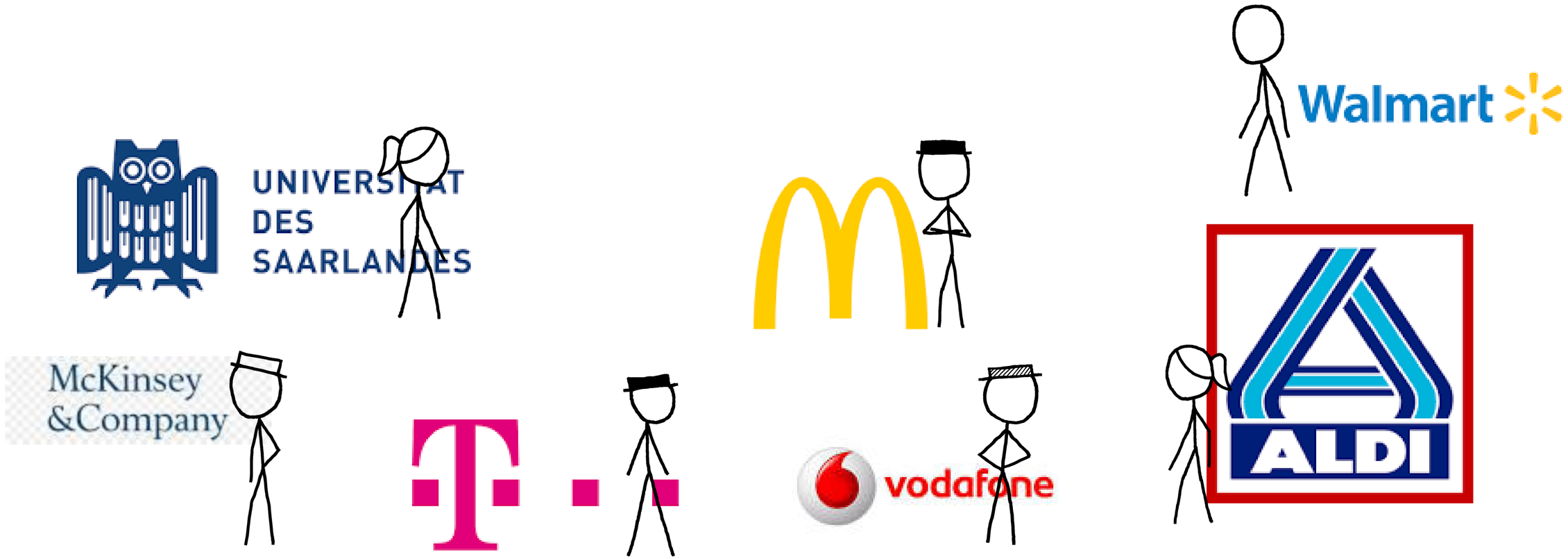
"[S]cale is what turns WMDs from local nuisances into tsunami forces, ones that define and delimit our lives." (from *Weapons of Math-Destruction* by Cathy O'Neil)



- Hopefully, there are better personnel managers at other companies out there…

WEAPONS OF MATH-DESTRUCTION: SCALE

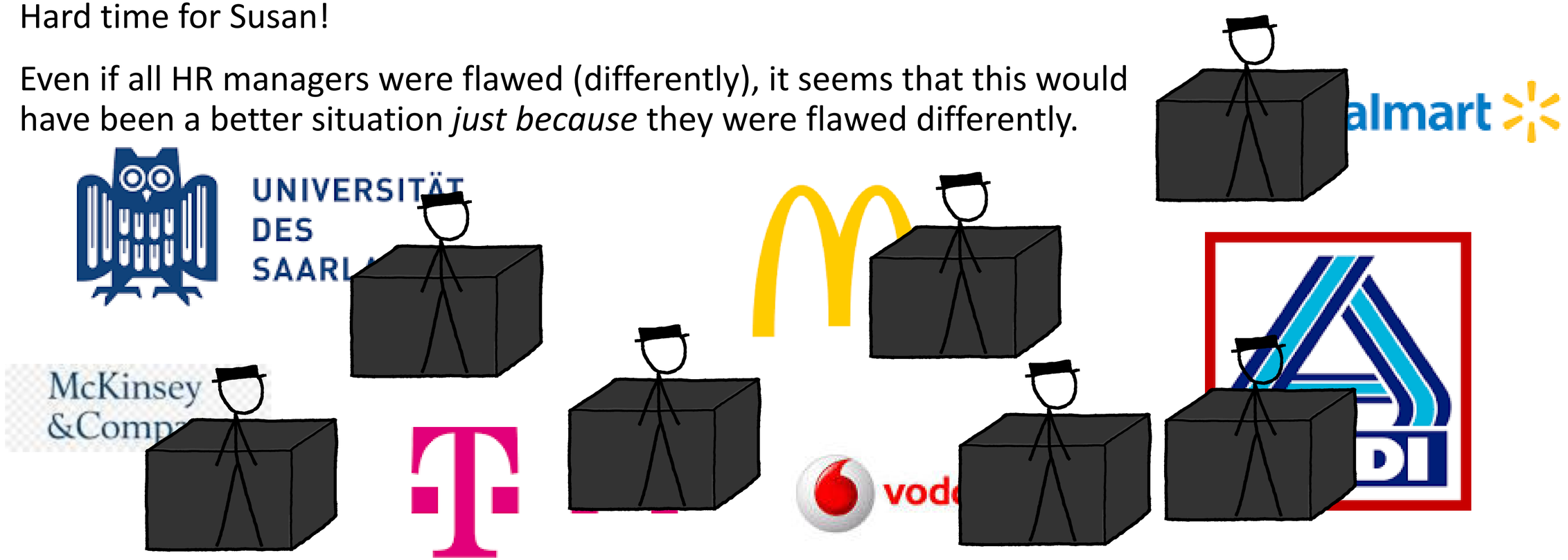
"[S]cale is what turns WMDs from local nuisances into tsunami forces, ones that define and delimit our lives."



WEAPONS OF MATH-DESTRUCTION: SCALE

"[S]cale is what turns WMDs from local nuisances into tsunami forces, ones that define and delimit our lives."

- Imagine, a recruiting recommendation AI based on Rudolph's decision.
- Imagine, further that this AI spreads to more and more businesses….
- Hard time for Susan!
- Even if all HR managers were flawed (differently), it seems that this would have been a better situation *just because* they were flawed differently.



WEAPONS OF MATH-DESTRUCTION: OPACITY

How do we know?

- Do people know what they are subject to ADMs or ASDMs?
- Should they know?
- What could they do against it?
- How to report mistakes?
- How to even sport mistakes if people do not know the inputs?
- And even if they knew the inputs, there might be mistakes impossible to spot in the weighing and reasoning if this is not known.

https://www.theguardian.com/science/2016/sep/01/how-algorithms-rule-our-working-lives

The long read

How algorithms rule our working lives

Employers are turning to mathematically modelled ways of sifting through job applications. Even when wrong, their verdicts seem beyond dispute – and they tend to punish the poor

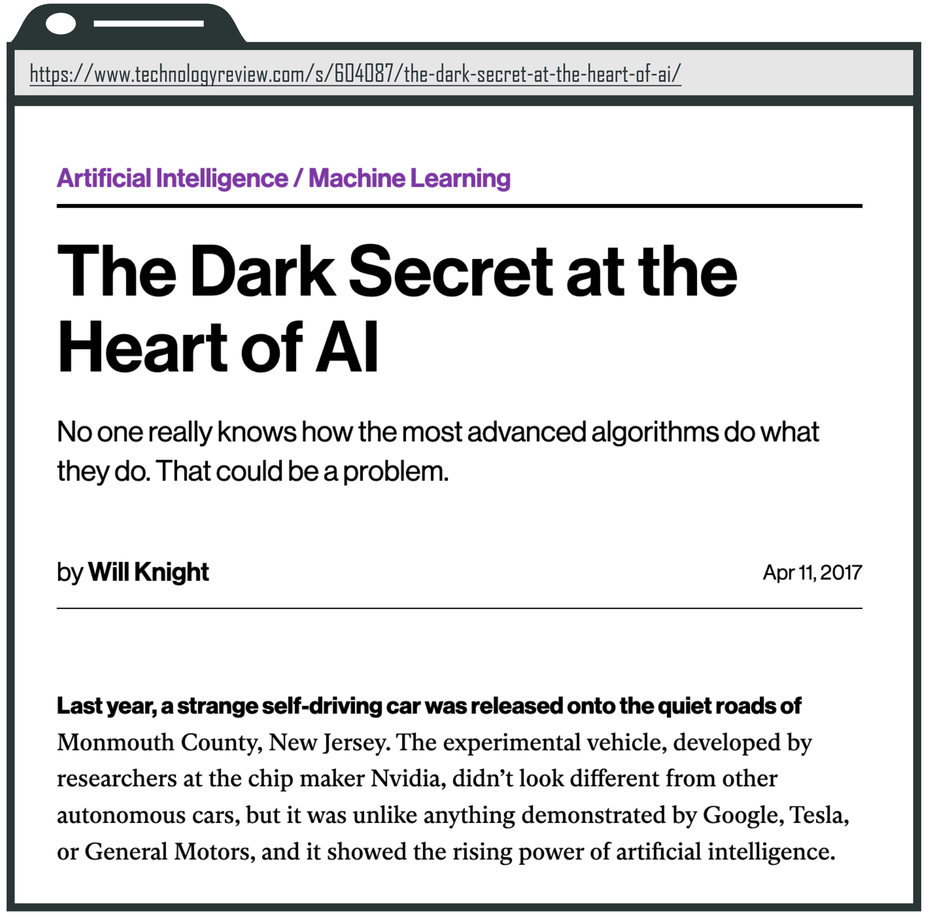
by Cathy O'Neil

A few years ago, a young man named Kyle Behm took a leave from his studies at Vanderbilt University in Nashville, Tennessee. He was suffering from bipolar disorder and needed time to get treatment. A year and a half later, Kyle was healthy enough to return to his studies at a different university. Around that time, he learned from a friend about a part-time job. It was just a minimum-wage job at a Kroger supermarket, but it seemed like a sure thing. His friend, who was leaving the job, could vouch for him. For a high-achieving student like Kyle, the application looked like a formality.

WEAPONS OF MATH-DESTRUCTION: OPACITY

How do we know?

- Do people know what they are subject to ADMs or ASDMs?
- Should they know?
- What could they do against it?
- How to report mistakes?
- How to even sport mistakes if people do not know the inputs?
- And even if they knew the inputs, there might be mistakes impossible to spot in the weighing and reasoning if this is not known.



https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Artificial Intelligence / Machine Learning

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

by **Will Knight** Apr 11, 2017

Last year, a strange self-driving car was released onto the quiet roads of Monmouth County, New Jersey. The experimental vehicle, developed by researchers at the chip maker Nvidia, didn't look different from other autonomous cars, but it was unlike anything demonstrated by Google, Tesla, or General Motors, and it showed the rising power of artificial intelligence.

WEAPONS OF MATH-DESTRUCTION: OPACITY

How do we know?

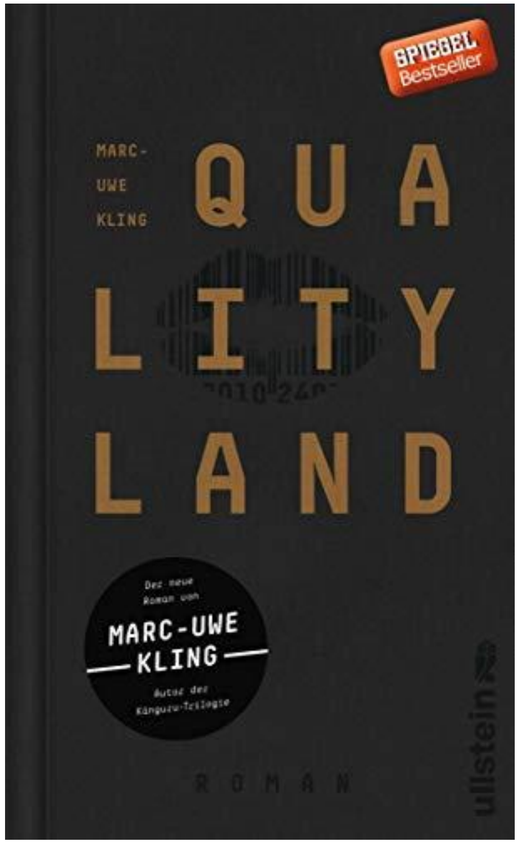
- Do people know what they are subject to ADMs or ASDMs?
- Should they know?
- What could they do against it?
- How to report mistakes?
- How to even sport mistakes if people do not know the inputs?
- And even if they knew the inputs, there might be mistakes impossible to spot in the weighing and reasoning if this is not known.
- How to know that reports have been considered?
- …

We need different kinds of accessibility, understandability, transparent.

In other words: Current models and their application/employment is too opaque!

Mark-Uwe Kling has coined the term “Peter’s Problem”:

The system tells me: I want that. But I do not!



TAKE-AWAY MESSAGE

Dismissing the “tech is just a tool” myth.

BEWARE THE MYTH: THE TOOL METAPHOR

Often you will hear something like:

> *“AI is just a tool. Whether it is beneficial or harmful depends on who it uses for what purpose.”*

This is true for the techniques we use: Coding, learning algorithms…

But what you code and the data you learn from has often already some values ‘impressed’ or purpose projected upon…

Furthermore, these are complex systems and they add up to complicated systems with many parts. These systems can have consequences no one intended and, in fact, no one was able to foresee.

Always be aware and keep that in mind:
Things are more complicated, you are not just toolmakers, you work on stuff that shapes societies and may decide over individuals’ futures.

https://www.frederickengraving.com/thumbnail.asp?file=assets/images/HM02.jpg&maxx=350&maxy=0

Ethics for Nerds

An Advanced Course in Computer Science
Summer Semester 2020

Current Topics C6.4
Algorithmic Decision-Making
& Algorithmically Supported Decision-Making

Machine Bias, Algorithmic Fairness, Discrimination

Prof. Holger Hermanns,
Kevin Baum, Sarah Sterz

UNIVERSITÄT
DES
SAARLANDES

FOCUS

General Problems

Embedded Values,
Math-Washing,
Pseudo-Solutions,
Self-Fulfilling Prophecies

Discrimination
& Algorithmic Fairness

Responsibility
& Explainability

(significantly shortened due to the current crisis)

FOCUS

General Problems

Embedded Values,
Math-Washing,
Pseudo-Solutions,
Self-Fulfilling Prophecies

Discrimination
& Algorithmic Fairness

Responsibility
& Explainability

(significantly shortened due to the current crisis)

Models & data can be (and, in fact, always will be) biased in different ways.

Know the limitations of your data and methods!

BIAS IS NOT THE SAME AS BIAS

“Bias” in social sciences etc.

(intuitively, roughly)

subjective or unjustified influenced judgement that makes it unfair or otherwise problematic

Connection?

“Bias” in statistics, data science, ML…

(intuitively, roughly)

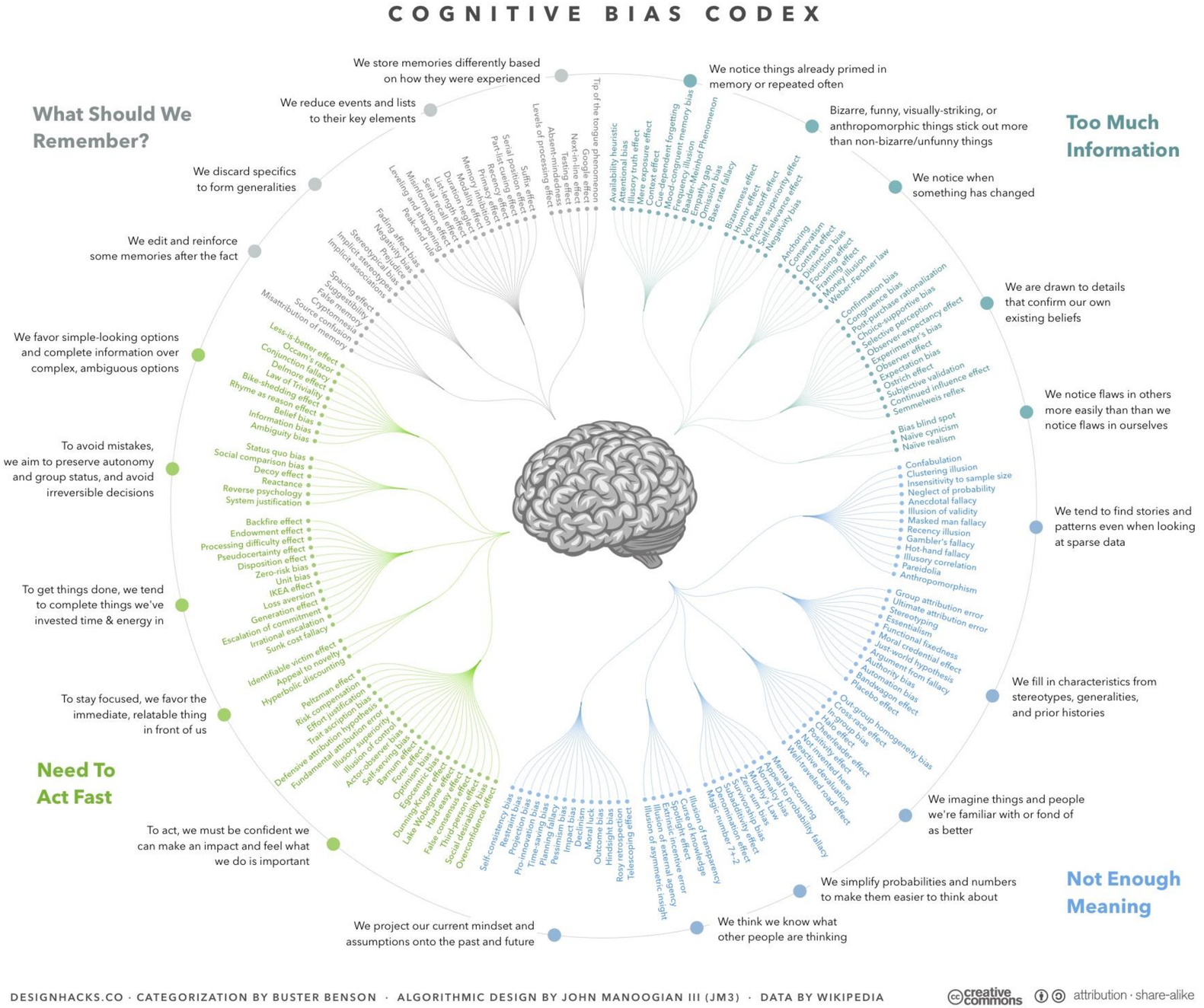
some data is not useful because of the data generating process; models are biased because they do not adequately represent reality, possibly because biased data was used.

- In many contexts it is not absolutely clear which of the meanings are meant.
- Sometimes the distinction is overlooked and also the connection between the two conceptions.

"Bias" in statistics, data science, ML...

(intuitively, roughly)

some data is not useful because of the data generating process; models are biased because they do not adequately represent reality, possibly because biased data was used.



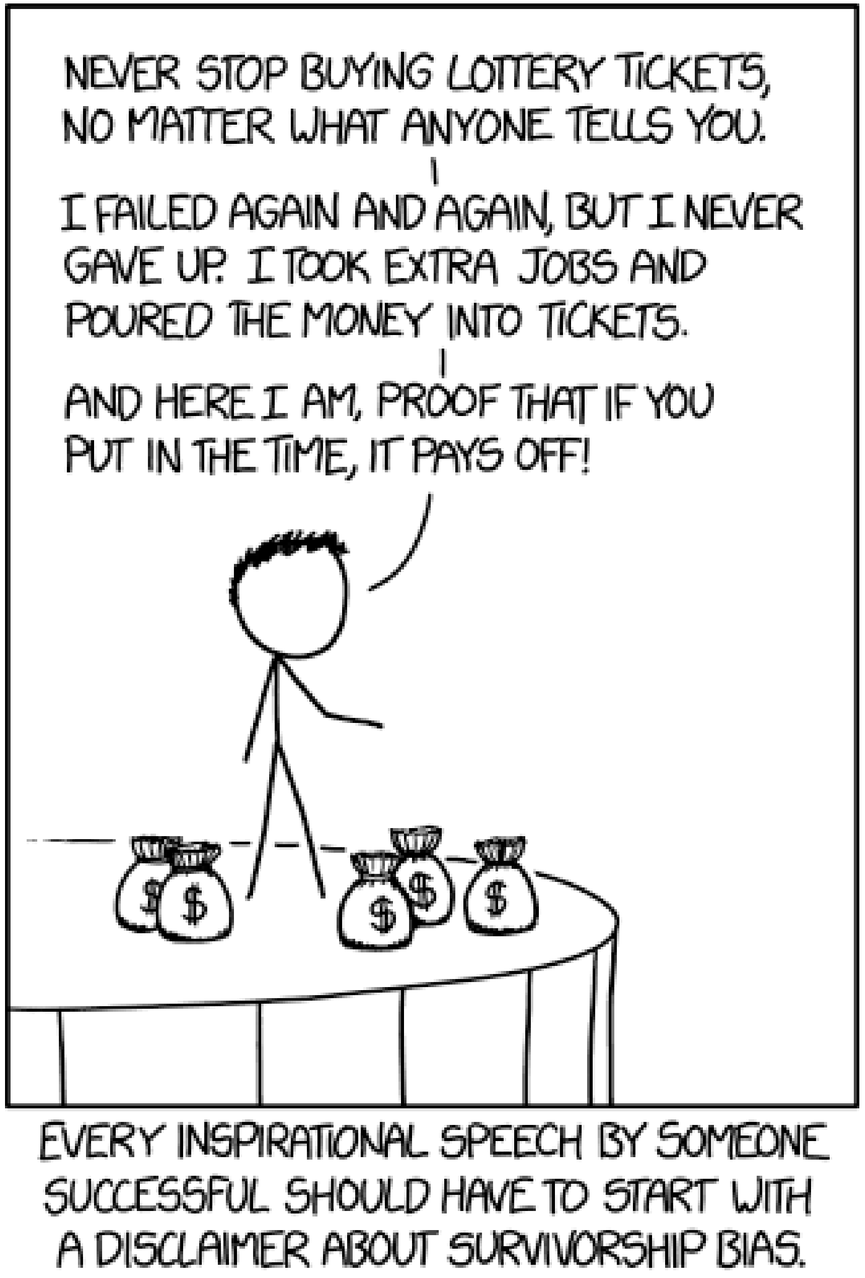
https://en.wikipedia.org/wiki/List_of_cognitive_biases
(unsurprisingly, it's quite long...)

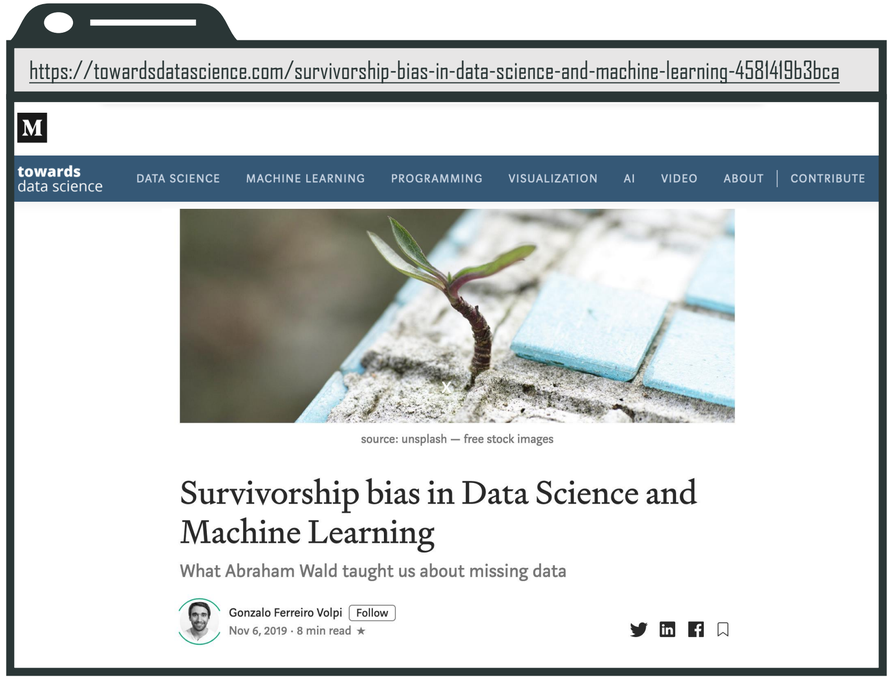
“Bias” in statistics, data science, ML…

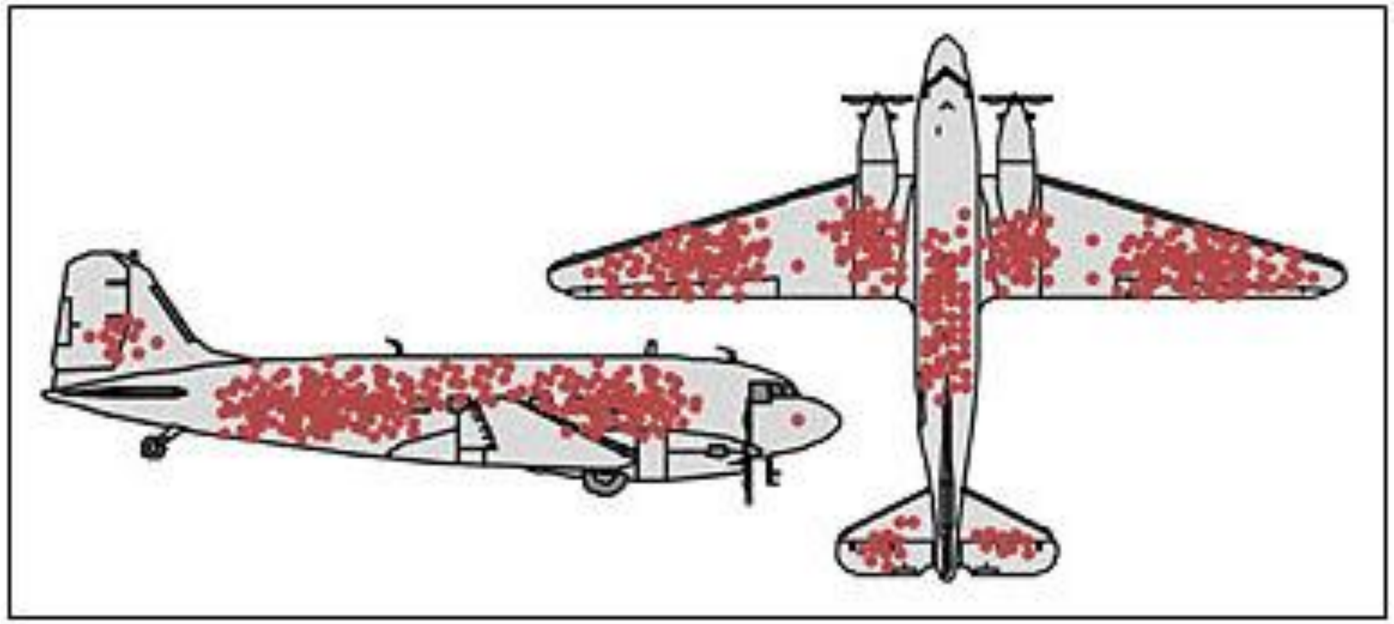
(intuitively, roughly)

some data is not useful because of the data generating process; models are biased because they do not adequately represent reality, possibly because biased data was used.

https://xkcd.com/1827/



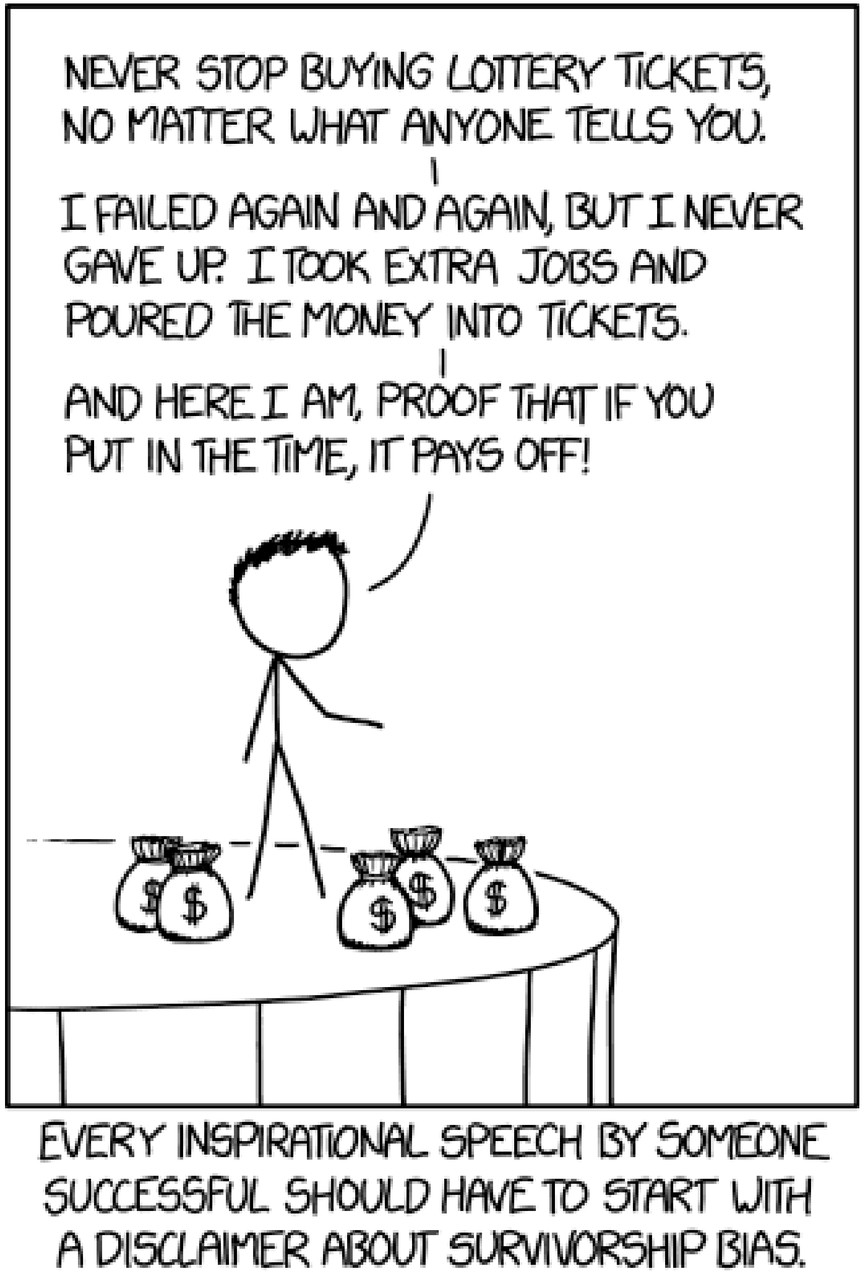




https://divakars.wordpress.com/2013/09/22/focusing-on-success-may-lead-to-survivorship-bias/

Credit: Cameron Moll

“Bias” in statistics, data science, ML…



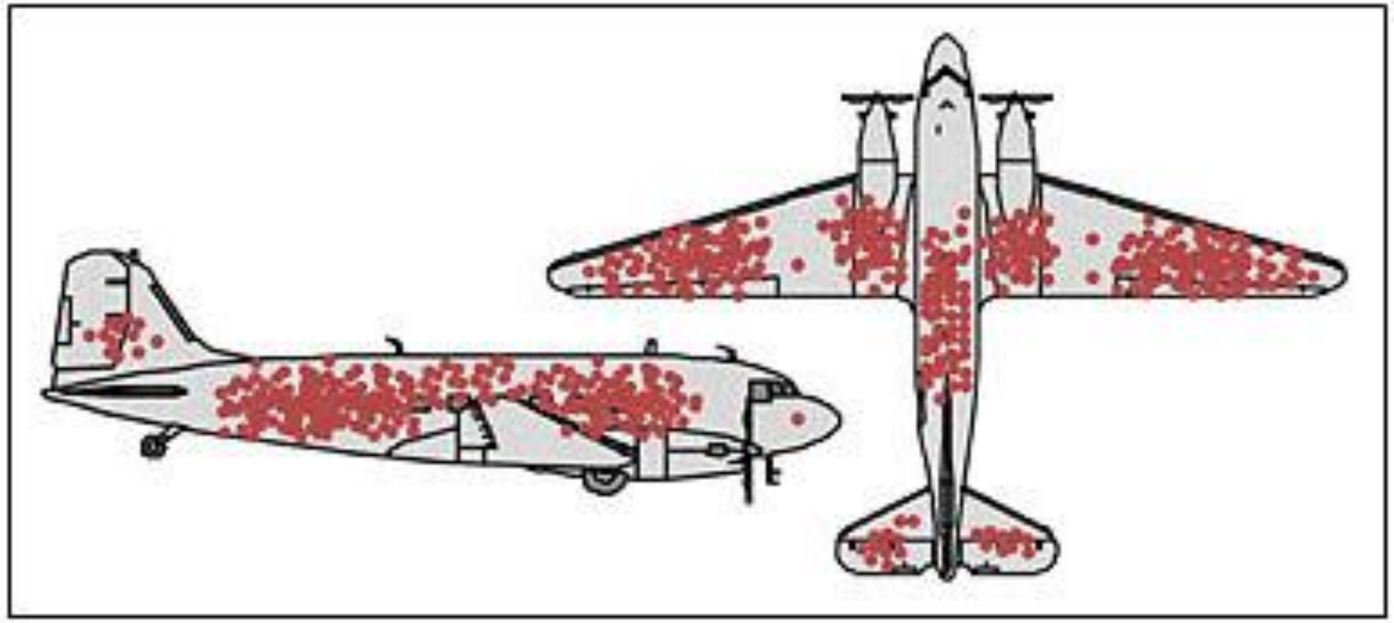
https://xkcd.com/1827/

https://towardsdatascience.com/survivorship-bias-in-data-science-and-machine-learning-4581419b3bca

- **Churn prediction**: suppose you’re building a machine learning model for churn prediction. So you take your current database of active customers and think about useful features you might use. The total number of orders? Number of orders in the last X months? Since when he has been with the company? The total number of complaints? The average number of interactions with customer service per month? Well, stuff like that. You think a lot and end up with the impressive number of 100 features for predicting churn…nice feature engineering you have done! Well done! However, have you already spotted the survivorship bias in action? Of course, taking your current active customers you would be missing all those who left you, and therefore, your model would only be looking at the survivors. The correct way of doing this would be using, for example, taking instead all customers who registered in a certain month of a certain year. In this way, you could achieve a representative sample of customers including those who are still with you and those who left.

(intuitively, roughly)

some data is not useful because of the data generating process; models are biased because they do not adequately represent reality, possibly because biased data was used.



https://divakars.wordpress.com/2013/09/22/focusing-on-success-may-lead-to-survivorship-bias/

Credit: Cameron Moll

BIASES IN ML

(intuitively, roughly)

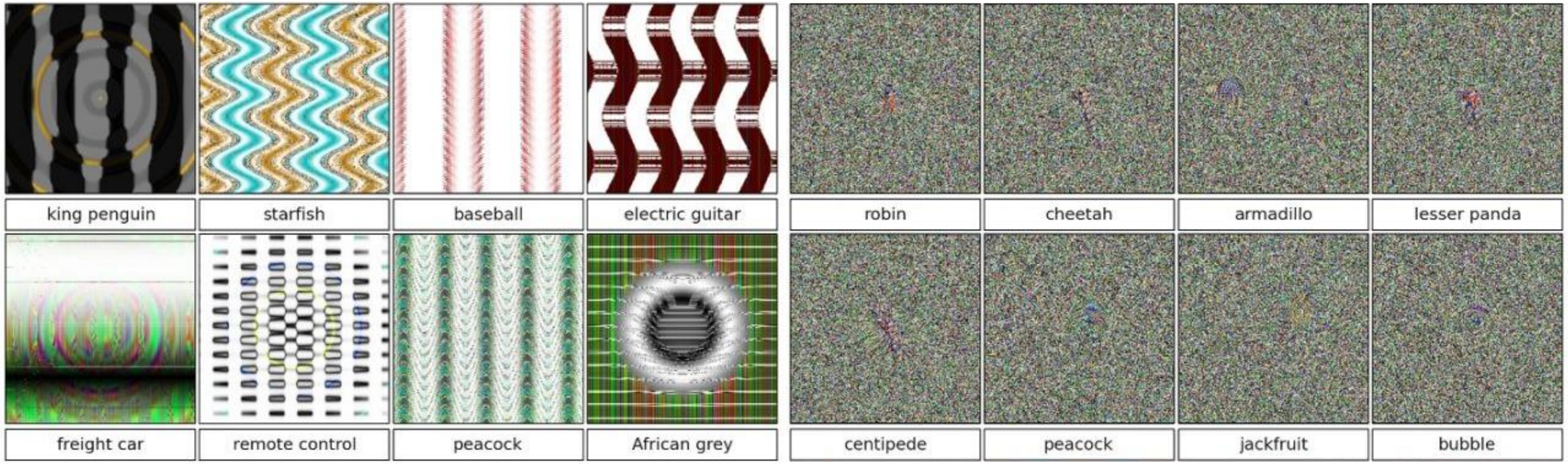
some data is not useful because of the data generating process; models are biased because they do not adequately represent reality, possibly because biased data was used.

I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.

Abraham Maslow

Applies in two ways:

- ML-algorithms seem to be the hammer, with which everything looks like a nail
  - which, ironically, is a bias for itself! “*Law of instrument*”
- ML-algorithms, see nails everywhere if they are (possibly implicitly) trained to do so



Nguyen et al: "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

BIASES IN ML

"Bias" in social sciences etc.

(intuitively, roughly)

subjective or unjustified influenced judgement that makes it unfair or otherwise problematic

Connection?

"Bias" in statistics, data science, ML...

(intuitively, roughly)

some data is not useful because of the data generating process; models are biased because they do not adequately represent reality, possibly because biased data was used.

Pierre Stock & Moustapha Cissé 2017: *ConvNets and ImageNet Beyond Accuracy: Explanations, Bias Detection, Adversarial Examples and Model Criticism*, https://arxiv.org/abs/1711.11443

COMPAS AS EXAMPLE FOR BIAS IN MODELS

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)

- COMPAS isn't a ML algorithm (but, for us, courts, judges and Loomis, a black box much like ML algos).
- We definitely know about COMPAS by now: it is deeply flawed! (We revisit this case later…)
- But is it *illegitimately* biased?

A Popular Algorithm Is No Better at Predicting Crimes Than Random People

The COMPAS tool is widely used to assess a defendant's risk of committing more crimes, but a new study puts its usefulness into perspective.

ED YONG | JAN 17, 2018 | TECHNOLOGY

https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

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

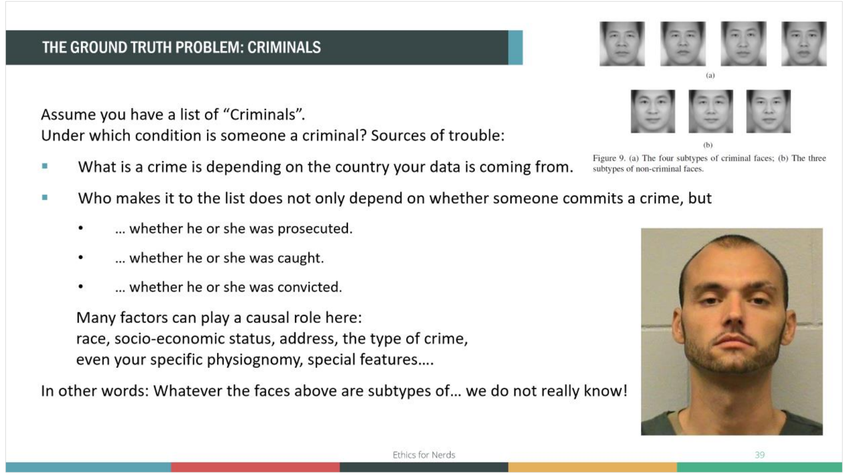
For a counter position see: https://medium.com/@AbeGong/ethics-for-powerful-algorithms-1-of-3-a060054efd84

AGAIN THE GROUND TRUTH PROBLEM: CRIMINALS

- A record of a criminal career is as record of interaction of *institutions* with *individuals*.
- Institutions consist of individuals.
- Individuals and institutions *can* be (and, in fact, *are*) biased.
- → Real world data often *will* be biased.
- Often there *will be no better* data available.

Worse: Sometimes it seems that there *cannot* be.

Even worse: Perfect fairness might be *impossible*, too…
(more on that in a second)

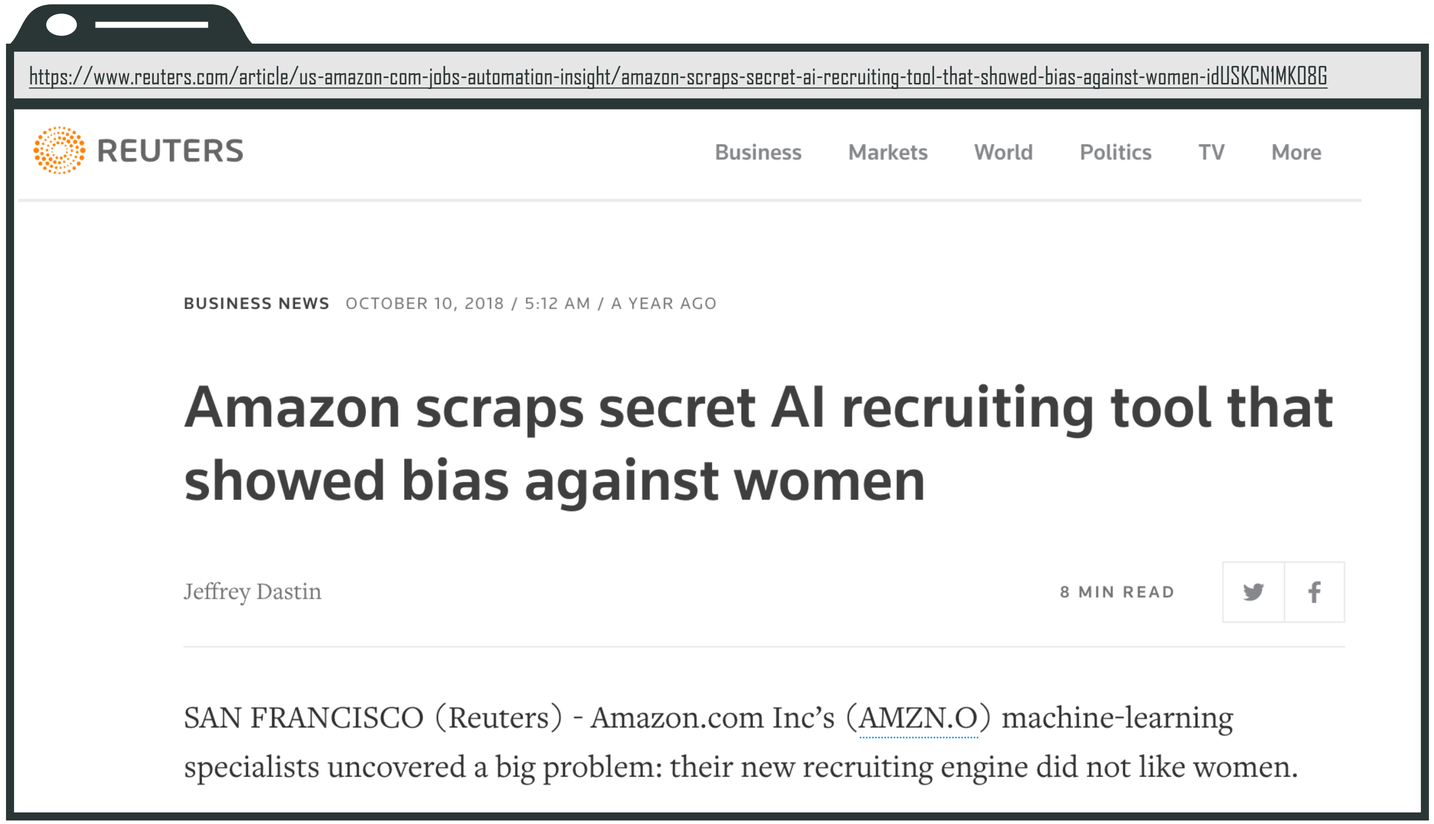


Not every bias is bad (i.e., morally problematic).

Not every bad bias is a case of unfair discrimination.

Some are… but what does it mean?

GENERAL IDEA


https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

REUTERS

Business Markets World Politics TV More

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Starting Point

Same treatment for qualitative alike cases; differential treatment for cases with qualitative differences.

There is a women's locker room and a men's locker room at school: different rooms for people of different gender.

GENERAL IDEA

Starting Point

Same treatment for qualitative alike cases; differential treatment for cases with qualitative differences.

There is a women's locker room and a men's locker room at school: different rooms for people of different gender.

Potential problem with this case: Is gender binary? What is about transgender people?

→ What are the relevant qualitative differences?

→ Depends on context!

PROTECTED ATTRIBUTES

→ Hot candidates for attributes/properties/features to handle with great care (default: they should deemed irrelevant!)

Starting Point

Same treatment for **qualitative alike** cases; differential treatment for cases with **qualitative differences**.

https://ohchr.org/EN/Issues/Discrimination/Pages/discrimination.aspx

- Combating discrimination against indigenous peoples
- Combating discrimination against migrants
- Combating discrimination against minorities
- Combating discrimination against people with disabilities
- Combating discrimination against women
- Combating racial discrimination
- Combating religious discrimination
- Combating discrimination based on sexual orientation and gender identity

1. *Charter of the United Nations*

PROTECTED ATTRIBUTES

But that is strictly speaking to narrow.

- Is it permissible to discriminate against people from some class (e.g. the poor)?
- Is it permissible to discriminate against names? (In Germany, “Kevin” is an example: https://en.wikipedia.org/wiki/Kevinism)
- Is it permissible to discriminate against someone because he is old (or young)?
- …

Call the set of properties/features/attributes which deserve special care **protected attributes**.

https://en.wikipedia.org/wiki/Homelessness#/media/File:HomelessParis_7032101.jpg

https://plato.stanford.edu/entries/discrimination/

WHAT IS ((UN-)FAIR) DISCRIMINATION?

By default we refer to the moralized concept hereafter.

Starting Point

Same treatment for qualitative alike cases; differential treatment for cases with qualitative differences.

Discrimination (morally neutral)

Acts, practices, or policies that impose a relative disadvantage on persons based on their membership in a salient social group.

Discrimination (moralized concept)

Acts, practices or policies that **wrongfully** impose a relative disadvantage on persons based on their membership in a salient social group of a suitable sort.

≙

Morally unjustified discrimination (in the neutral sense).

Be careful: depending on what concept is used, the sentence “this discrimination is wrong” is either a tautology or a substantive moral claim! (In many debates and articles these concepts get mixed up…)

https://plato.stanford.edu/entries/discrimination/

DIRECTS & INDIRECT DISCRIMINATION

Direct Discrimination

Discrimination grounded in some objectionable mental state.

Examples:

- **Direct reference:**
  “No Muslims are allowed at our school!”
- **Intended discrimination:**
  “Applications only on Saturdays!”
  (for religious reasons, orthodox Jews are not allowed to do any work on Saturdays)
- **Unwarranted beliefs:** “I help women! I hire them wherever I can, but obviously only for the kind of jobs they are capable of. Certainly not software development!”

Indirect Discrimination

“[w]hen a general policy or measure has disproportionately prejudicial effects on a particular group, it is not excluded that this may be considered as discriminatory notwithstanding that it is not specifically aimed or directed at that group” (ECHR)

[A] policy with disproportionate effects counts as indirect discrimination “if it does not pursue a legitimate aim or if there is not a reasonable relation of proportionality between means and aim” (ECHR)

[A] policy with disproportionate effects is discriminatory “if it is not based on objective and reasonable criteria” (Human Rights Commitee of the UN)

Example: Using a written test to determine promotions for a job where writing skills are irrelevant and in an area where black employees have significant worse writing skills (which is unbeknown to the employer).

https://plato.stanford.edu/entries/discrimination/

DIRECTS & INDIRECT DISCRIMINATION

For the context of this lecture:

Direct Discrimination

Discrimination grounded in some objectionable mental state.

Examples:

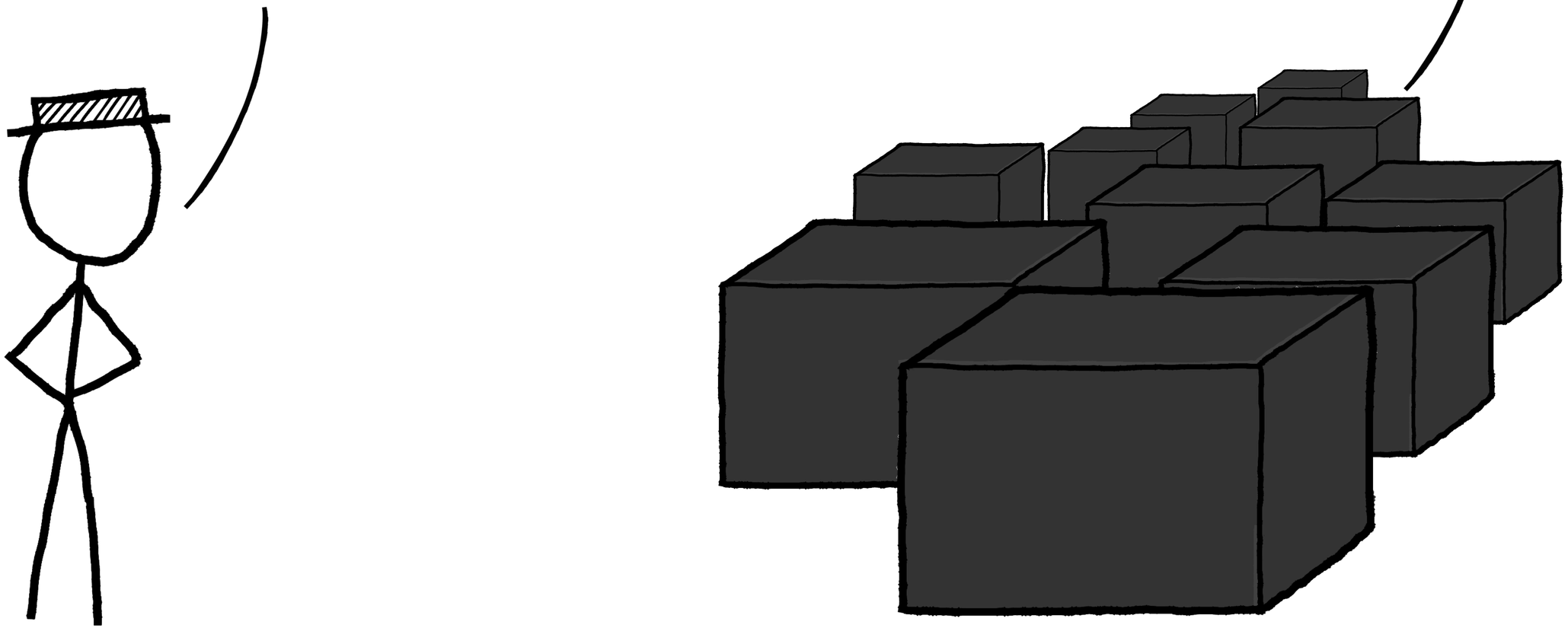
- **Direct reference:**
  "No Muslims are allowed at our school!"
- **Intended discrimination:**
  "Applications only on Saturdays!"
  (for religious reasons, orthodox Jews are not allowed to do any work on Saturdays)
- **Unwarranted beliefs:** "I help women! I hire them wherever I can, but obviously only for the kind of jobs they are capable of. Certainly not software development!"

Indirect Discrimination

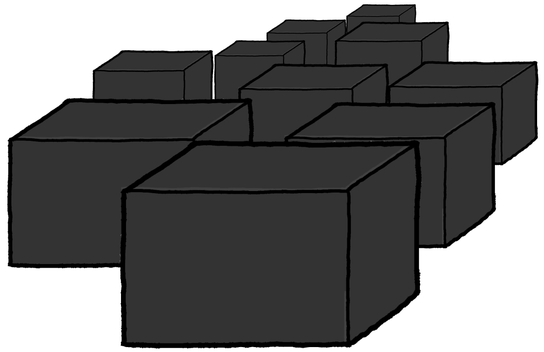
[A] policy with disproportionate effects is discriminatory "if it is not based on objective and reasonable criteria" (Human Rights Commitee of the UN)

Example: Using a written test to determine promotions for a job where writing skills are irrelevant and in an area where black employees have significant worse writing skills (which is unbeknown to the employer).

CAN ALGORITHMS DISCRIMINATE?



TECHNOLOGY CAN DISCRIMINATE!



1. direct discrimination grounded in “own” mental states ✗
2. direct discrimination grounded in developer’s mental states ✓
3. indirect discrimination ✓

https://www.mic.com/articles/124899/the-reason-this-racist-soap-dispenser-doesn-t-work-on-black-skin

The Reason This "Racist Soap Dispenser" Doesn't Work on Black Skin

By **Max Plenke** | Sept. 9, 2015

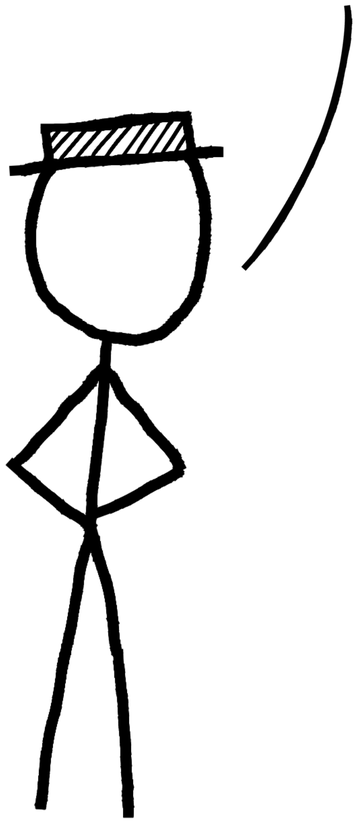
At a Marriott hotel in Atlanta, the soap dispensers have a little bit of a race problem.

Fitzpatrick made it, and it's clear this is a joke to him. But it introduces the more pervasive problem of technology being constructed without paying mind to the diversity of bodies it is built to serve.

What to do against discrimination through technology?

Welcome to the field of *algorithmic fairness*!

RESPONSES TO THE POSSIBILITY OF DISCRIMINATING MODELLS

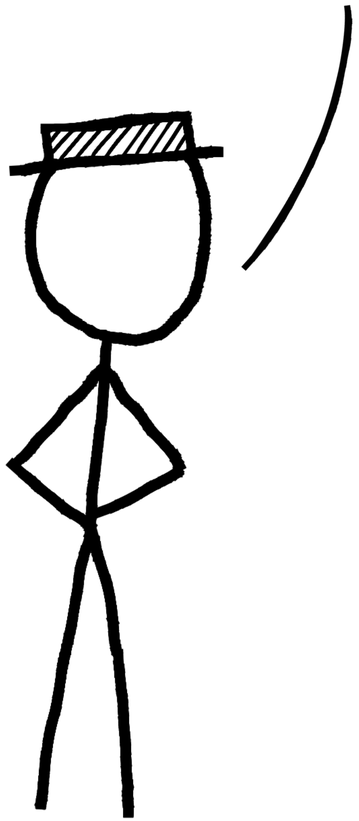


Okay, but if my model has no bias at all, then it is useless!

Answer:

Yeah, you mix up discrimination in the moral neutral and the moralized sense. You mean statistical bias. That‘s okay, as long as your model is not discriminative regarding protected attributes!

RESPONSES TO THE POSSIBILITY OF DISCRIMINATING MODELLS



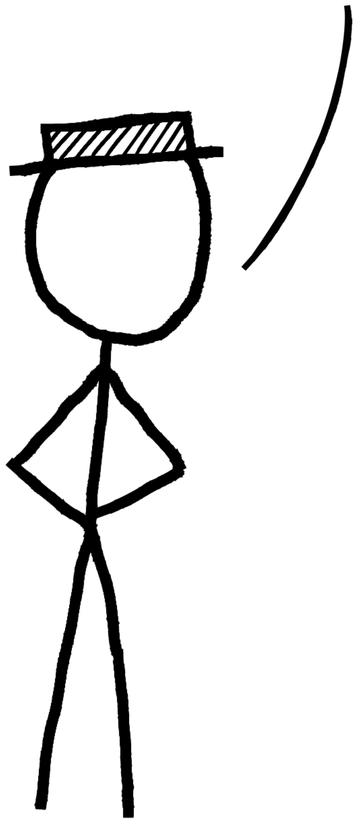
Okay, I just do not give my model access to protected attributes!

Answer:

Nice try, but this won't do the job. Other attributes might correlate with protected attributes. Therefore your model still might discriminate against protected attributes!

Direct references is not necessary for discrimination (not even for direct discrimination)!

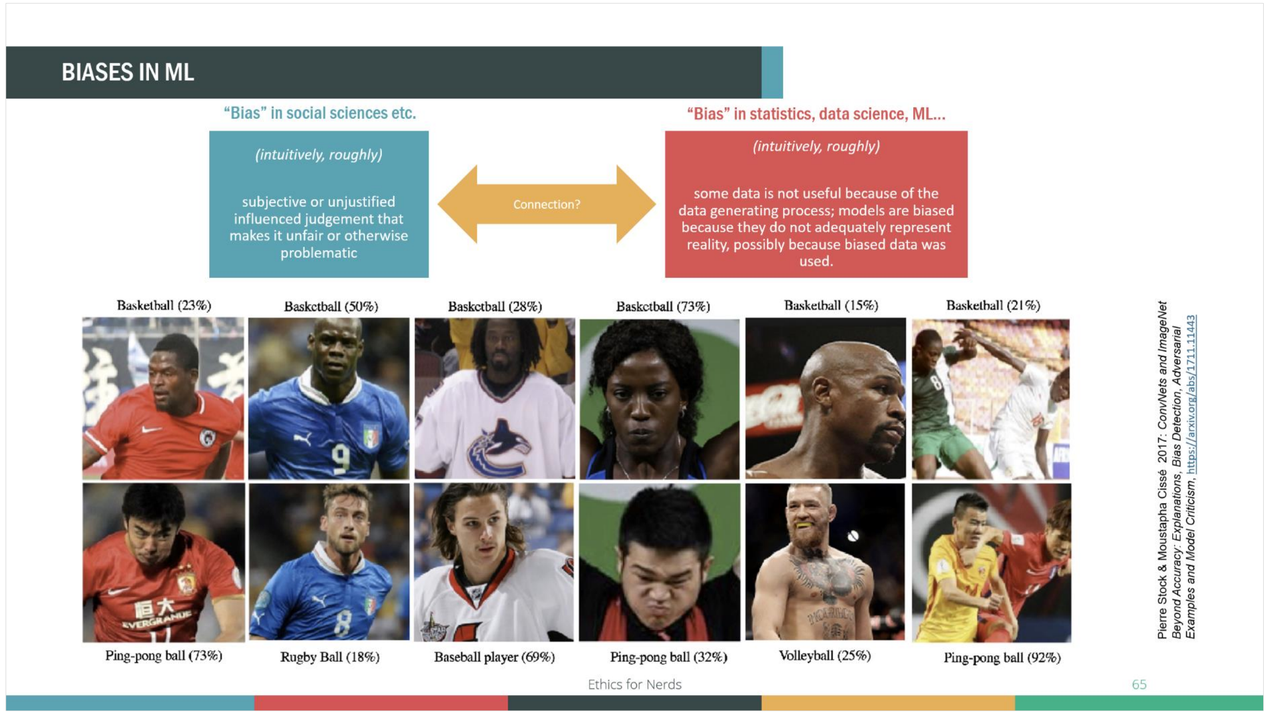
RESPONSES TO THE POSSIBILITY OF DISCRIMINATING MODELLS



Okay, wait a moment… what do you even mean by relative disadvantage? What are we talking about? Sensitivity or specificity or …?

Answer:

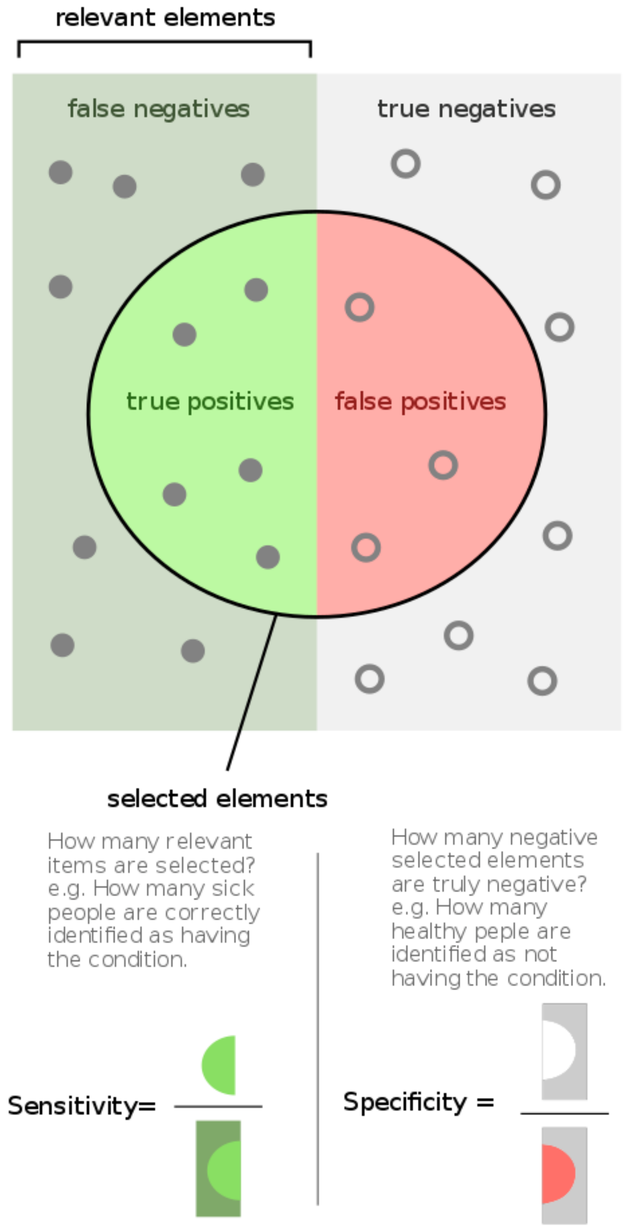
Well… wait, what?



DIFFERENT TREATMENT REGARDING...WHAT?

		Predicted class: P	Predicted class: N
Actual class	P	TP	FN
	N	FP	TN

		True condition: Condition positive	True condition: Condition negative		
	Total population	Condition positive	Condition negative	Prevalence = Σ Condition positive / Σ Total population	Accuracy (ACC) = Σ True positive + Σ True negative / Σ Total population
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = Σ True positive / Σ Predicted condition positive	False discovery rate (FDR) = Σ False positive / Σ Predicted condition positive
Predicted condition	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = Σ False negative / Σ Predicted condition negative	Negative predictive value (NPV) = Σ True negative / Σ Predicted condition negative
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = Σ True positive / Σ Condition positive	False positive rate (FPR), Fall-out, probability of false alarm = Σ False positive / Σ Condition negative	Positive likelihood ratio (LR+) = TPR / FPR	Diagnostic odds ratio (DOR) = LR+ / LR−
		False negative rate (FNR), Miss rate = Σ False negative / Σ Condition positive	Specificity (SPC), Selectivity, True negative rate (TNR) = Σ True negative / Σ Condition negative	Negative likelihood ratio (LR−) = FNR / TNR	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$



Terminology and derivations from a confusion matrix

condition positive (P)
the number of real positive cases in the data

condition negative (N)
the number of real negative cases in the data

true positive (TP)
eqv. with hit

true negative (TN)
eqv. with correct rejection

false positive (FP)
eqv. with false alarm, Type I error

false negative (FN)
eqv. with miss, Type II error

sensitivity, recall, hit rate, or true positive rate (TPR)
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

specificity, selectivity or true negative rate (TNR)
$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$

precision or positive predictive value (PPV)
$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{FDR}$$

negative predictive value (NPV)
$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} = 1 - \mathrm{FOR}$$

miss rate or false negative rate (FNR)
$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR}$$

fall-out or false positive rate (FPR)
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

false discovery rate (FDR)
$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{PPV}$$

false omission rate (FOR)
$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{NPV}$$

Prevalence Threshold (PT)
$$PT = \frac{\sqrt{TPR(-TNR+1)} + TNR - 1}{(TPR + TNR - 1)}$$

Threat score (TS) or critical success index (CSI)
$$\mathrm{TS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

accuracy (ACC)
$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

balanced accuracy (BA)
$$\mathrm{BA} = \frac{TPR + TNR}{2}$$

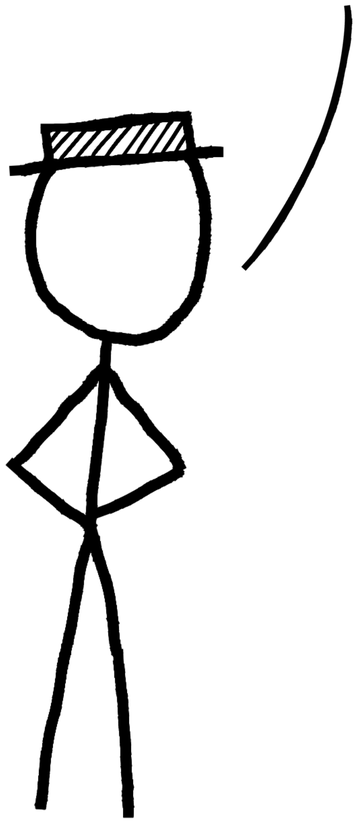
F1 score
is the harmonic mean of precision and sensitivity
$$\mathrm{F_1} = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

https://en.wikipedia.org/wiki/Confusion_matrix

RESPONSES TO THE POSSIBILITY OF DISCRIMINATING MODELLS

Let me check …



Answer:

Well… wait, what?

We never thought about that in such formal terms…

Can’t you calibrate for all these possible errors?

FAIRNESS TO THE RESCUE? (Statistical Fairness)

https://arxiv.org/pdf/1703.00056.pdf

https://www.liebertpub.com/doi/abs/10.1089/big.2016.0047

Fair prediction with disparate impact:

A study of bias in recidivism prediction instruments

Alexandra Chouldechova

Heinz College, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA, USA
achould@cmu.edu

Let $S = S(x)$ denote the risk score based on covariates $X = x$, with higher values of S corresponding to higher levels of assessed risk. Let $R \in \{b, w\}$ denote the group that the individual belongs to, which may be one of the components of X. Lastly, let $Y \in \{0, 1\}$ be the outcome indicator, with 1 denoting that the given individual recidivates.

Definition 1 (Calibration)**.** A score $S = S(x)$ is said to be *well-calibrated* if it reflects the same likelihood of recidivism irrespective of the individuals' group membership. That is, if for all values of s,

$$\mathbb{P}(Y = 1 \mid S = s, R = b) = \mathbb{P}(Y = 1 \mid S = s, R = w). \tag{2.1}$$

Definition 2 (Predictive parity)**.** A score $S = S(x)$ satisfies *predictive parity* at a threshold s_{HR} if the likelihood of recidivism among high-risk offenders is the same regardless of group membership. That is, if,

$$\mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = b) = \mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = w). \tag{2.2}$$

Definition 3 (Error rate balance)**.** A score $S = S(x)$ satisfies *error rate balance* at a threshold s_{HR} if the false positive and false negative error rates are equal across groups. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = b) = \mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = w)\,, \quad \text{and} \tag{2.3}$$
$$\mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = b) = \mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = w), \tag{2.4}$$

where the expressions in the first line are the group-specific false positive rates, and those in the second line are the group-specific false negative rates.

Definition 4 (Statistical parity)**.** A score $S = S(x)$ satisfies *statistical parity* at a threshold s_{HR} if the proportion of individuals classified as high-risk is the same for each group. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid R = b) = \mathbb{P}(S > s_{\text{HR}} \mid R = w) \tag{2.5}$$

Statistical parity also goes by the name of *equal acceptance rates*[14] or *group fairness*[15],

FAIRNESS TO THE RESCUE? (Statistical Fairness)

https://arxiv.org/pdf/1703.00056.pdf
https://www.liebertpub.com/doi/abs/10.1089/big.2016.0047

All of the fairness metrics presented in the Background section can be thought of as imposing constraints on the values (or the distribution of values) in this table. Another constraint—one that we have no direct control over—is imposed by the recidivism prevalence within groups. It is not difficult to show that the prevalence (p), PPV, and false positive and negative error rates (FPR, FNR) are related through the equation

$$\mathrm{FPR} = \frac{p}{1-p}\frac{1-\mathrm{PPV}}{\mathrm{PPV}}(1-\mathrm{FNR}). \qquad (2.6)$$

From this simple expression, we can see that if an instrument satisfies predictive parity—that is, if the PPV is the same across groups—but the prevalence differs between groups, the instrument cannot achieve equal FPRs and FNRs across those groups.

FAIRNESS TO THE RESCUE? (Statistical Fairness)

$$\mathrm{FPR} = \frac{p}{1-p}\frac{1-\mathrm{PPV}}{\mathrm{PPV}}(1-\mathrm{FNR}).$$

(1) $FPR_B = \frac{p_B}{1-p_B}\frac{1-PPV_B}{PPV_B}(1-FNR_B)$

(2) $FPR_W = \frac{p_W}{1-p_W}\frac{1-PPV_W}{PPV_W}(1-FNR_W)$

Assuming that $PPV_B = PPV_W, FNR_B = FNR_W$ we can refomulate

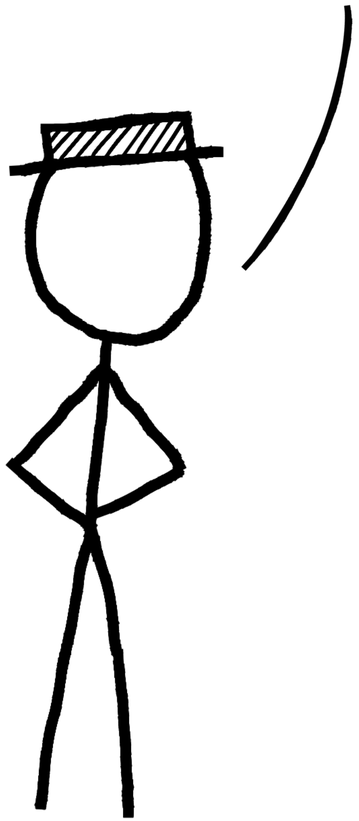
(3) $FPR_B = \frac{p_B}{1-p_B}\frac{1-PPV_W}{PPV_W}(1-FNR_W)$

Assume now that $FPR_B = FPR_W$

(4) $\frac{p_B}{1-p_B}\frac{1-PPV_W}{PPV_W}(1-FNR_W) =^{?} \frac{p_W}{1-p_W}\frac{1-PPV_W}{PPV_W}(1-FNR_W) \iff \frac{p_B}{1-p_B} =^{?} \frac{p_W}{1-p_W}$

$\iff p_B =^{?} p_w$ which, by assumption is not the case. ■

RESPONSES TO THE POSSIBILITY OF DISCRIMINATING MODELLS



To be honest: No! Mathematically impossible.

But we could find the *right* balance and trade-offs between them!

Answer:

Well… wait, what?

We never thought about that in such formal terms…

Can't you calibrate for all these possible errors?

Well, the right measure depends on the context… You will need ethicists, sociologists, psychologists, and experts from the law faculty to identify appropriate measures…

Let me explain:

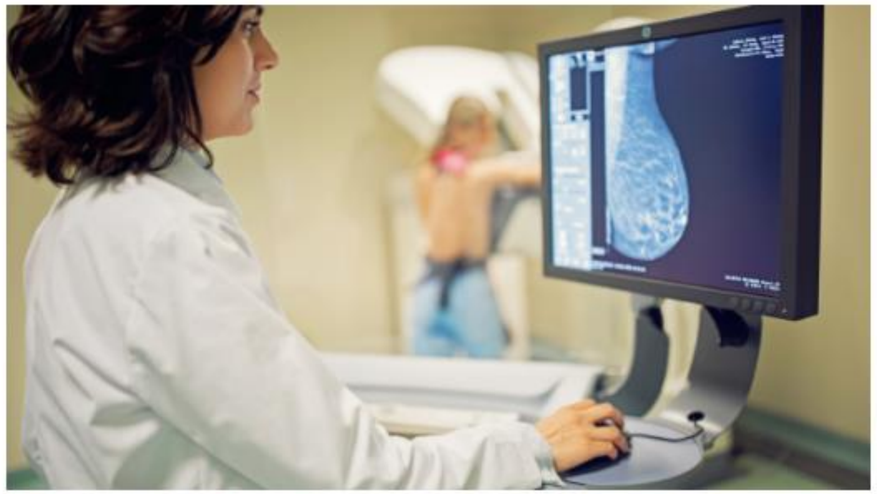
PREVENTATION AND EARLY DIAGNOSTICS

https://www.technologyreview.com/2020/01/03/238154/googles-ai-breast-cancer-screening-tool-is-learning-to-generalize-across-countries/

MIT Technology Review

Artificial intelligence Jan 03

Google's AI breast cancer screening tool is learning to generalize across countries



In a preliminary test, a model trained only on data from UK women still performed better than experts on US patients.

The news: DeepMind and Google Health have developed a new AI system to help doctors detect breast cancer early. The researchers trained an algorithm on mammogram images from female patients in the US and UK, and it performed better than human radiologists. The results were published in Nature on Wednesday.

A tragedy of errors: Breast cancer is the most common cancer for women globally, and their second leading cause of death. Though early detection and treatment can improve a patient's prognosis, screening tests have high rates of error. About 1 in 5 screenings fail to find breast cancer even when it's present, also known as a false negative; 50% of women who receive annual mammograms also get at least one false alarm over a 10-year period, known as a false positive.

The results: In tests, the AI system decreased both types of error. For US patients, it reduced false negatives and positives by 9.4% and 5.7%, respectively; for UK patients it reduced them by 2.7% and 1.2%. In a separate experiment, the researchers tested the system's ability to generalize: they trained the model using only mammograms from UK patients, and then evaluated its performance on US patients. The system still outperformed human radiologists, reducing false negatives and positives by 8.1% and 3.5%.

Imagine:

More false positives for women older than 50.

Compare to:

More false negatives for women older than 50.

What's to prefer?

(false positives!)

RISK ASSESSMENT IN TRIALS

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

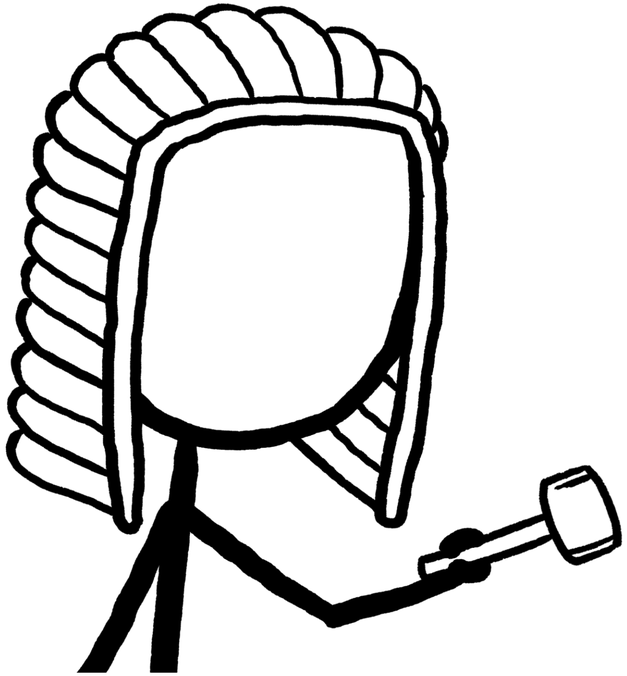
Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016



Imagine:

More false positives for people of color.

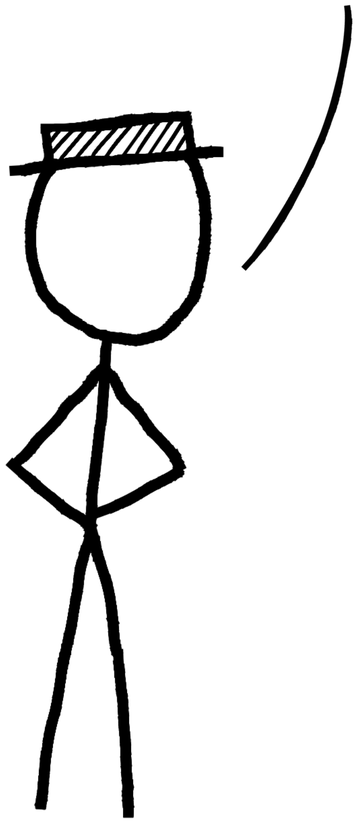
Compare to:

More false negatives for people of color.

What's to prefer?

(false negatives, at least according to a liberal and post-enlightenment society!)

RESPONSES TO THE POSSIBILITY OF DISCRIMINATING MODELLS



Okay, but then tell me in which context to calibrate for what error!

Answer:

Well… wait, what?

We never thought about that in such formal terms…

Can't you calibrate for all these possible errors?

Well, the right measure depends on the context… You will need ethicists, sociologists, psychologists, and experts from the law faculty to identify appropriate measures…

Let me explain…

→ Research needed!

→ Regulation needed!

RESPONSES TO THE POSSIBILITY OF DISCRIMINATING MODELLS

Answer:

How is that?

FAIRNESS TO THE RESCUE? (COUNTERFACTUAL FAIRNESS WITH CAUSAL MODELS)

https://papers.nips.cc/paper/6995-counterfactual-fairness.pdf

Counterfactual Fairness

Matt Kusner *
The Alan Turing Institute and
University of Warwick
mkusner@turing.ac.uk

Joshua Loftus *
New York University
loftus@nyu.edu

Chris Russell *
The Alan Turing Institute and
University of Surrey
crussell@turing.ac.uk

Ricardo Silva
The Alan Turing Institute and
University College London
ricardo@stats.ucl.ac.uk

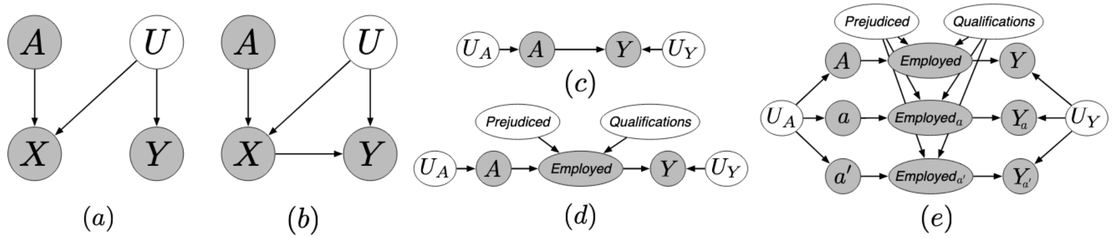


Figure 1: (a), (b) Two causal models for different real-world fair prediction scenarios. See Section 3.1 for discussion. (c) The graph corresponding to a causal model with A being the protected attribute and Y some outcome of interest, with background variables assumed to be independent. (d) Expanding the model to include an intermediate variable indicating whether the individual is employed with two (latent) background variables **Prejudiced** (if the person offering the job is prejudiced) and **Qualifications** (a measure of the individual's qualifications). (e) A twin network representation of this system [28] under two different counterfactual levels for A. This is created by copying nodes descending from A, which inherit unaffected parents from the factual world.

We perform inference on this model using an observed training set to estimate the posterior distribution of K. We use the probabilistic programming language Stan [34] to learn K. We call the predictor constructed using K, **Fair** K.

Fair prediction. As described in Section 4.2, there are three ways in which we can model a counterfactually fair predictor of FYA. Level 1 uses any features which are not descendants of race and sex for prediction. Level 2 models latent 'fair' variables which are parents of observed variables. These variables are independent of both race and sex. Level 3 models the data using an additive error model, and uses the independent error terms to make predictions. These models make increasingly strong assumptions corresponding to increased predictive power. We split the dataset 80/20 into a train/test set, preserving label balance, to evaluate the models.

As we believe LSAT, GPA, and FYA are all biased by race and sex, we cannot use any observed features to construct a counterfactually fair predictor as described in Level 1.

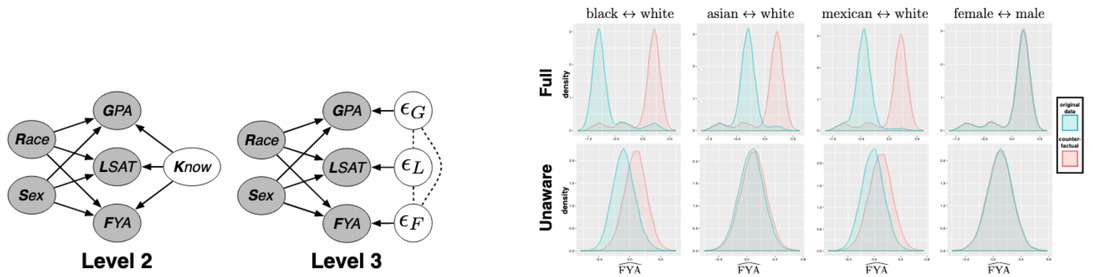


Figure 2: **Left:** A causal model for the problem of predicting law school success fairly. **Right:** Density plots of predicted FYA_a and $\text{FYA}_{a'}$.

Ethics for Nerds

An Advanced Course in Computer Science
Summer Semester 2020

Current Topics C6.5
Algorithmic Decision-Making
& Algorithmically Supported Decision-Making

Responsibility & Explainability

Prof. Holger Hermanns,
Kevin Baum, Sarah Sterz

UNIVERSITÄT
DES
SAARLANDES

FOCUS

General Problems

Embedded Values,
Math-Washing,
Pseudo-Solutions,
Self-Fulfilling Prophecies

Discrimination
& Algorithmic Fairness

Responsibility
& Explainability

(significantly shortened due to the current crisis)

FOCUS

General Problems

Embedded Values,
Math-Washing,
Pseudo-Solutions,
Self-Fulfilling Prophecies

Discrimination
& Algorithmic Fairness

Responsibility
& Explainability

(significantly shortened due to the current crisis)

https://link.springer.com/article/10.1007/s10676-004-3422-1

Published: September 2004

The responsibility gap: Ascribing responsibility for the actions of learning automata

Andreas Matthias

Ethics and Information Technology **6**, 175–183(2004) | Cite this article

Abstract

Traditionally, the manufacturer/operator of a machine is held (morally and legally) responsible for the consequences of its operation. Autonomous, learning machines, based on neural networks, genetic algorithms and agent architectures, create a new situation, where the manufacturer/operator of the machine is *in principle* not capable of predicting the future machine behaviour any more, and thus cannot be held morally responsible or liable for it. The society must decide between not using this kind of machine any more (which is not a realistic option), or facing a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3531711

Artificial Intelligence in Finance: Putting the Human in the Loop

CFTE Academic Paper Series: Centre for Finance, Technology and Entrepreneurship, no. 1.

50 Pages • Posted: 3 Mar 2020 • Last revised: 25 Mar 2020

Dirk A. Zetzsche
Universite du Luxembourg - Faculty of Law, Economics and Finance; Heinrich Heine University Dusseldorf - Center for Business & Corporate Law (CBC)

Douglas W. Arner
The University of Hong Kong - Faculty of Law

Ross P. Buckley
University of New South Wales (UNSW) - Faculty of Law

Brian Tang
The University of Hong Kong - Faculty of Law

Date Written: February 1, 2020

Abstract

Finance has become one of the most globalized and digitized sectors of the economy. It is also one of the most regulated of sectors, especially since the 2008 Global Financial Crisis. Globalization, digitization and money are propelling AI in finance forward at an ever increasing pace.

This paper develops a regulatory roadmap for understanding and addressing the increasing role of AI in finance, focusing on human responsibility: the idea of "putting the human in the loop" in order in particular to address "black box" issues.

https://futureoflife.org/2017/11/30/human-control-principle

When Should Machines Make Decisions?

November 30, 2017 / by Ariel Conn

Click here to see this page in other languages: Chinese Russian

Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

When is it okay to let a machine make a decision instead of a person? Most of us allow Google Maps to choose the best route to a new location. Many of us are excited to let self-driving cars take us to our destinations while we work or

[...]

[...]

In addition to the issues already mentioned with decision-making machines, Patrick Lin, a philosopher at California Polytechnic State University, doesn't believe it's clear who would be held responsible if something does go wrong.

"I wouldn't say that you must always have meaningful human control in everything you do," Lin said. "I mean, it depends on the decision, but also I think this gives rise to new challenges. ... This is related to the idea of human control and responsibility. If you don't have human control, it could be unclear who's responsible ... the context matters. It really does depend on what kind of decisions we're talking about, that will help determine how much human control there needs to be."

[...]

What is the role of a human decision-maker?

FLAVORS OF (MACHINE) GUIDED DECISIONS

Level 1

Full human autonomy

Level 2

Remote control

Level 3

Human in the loop

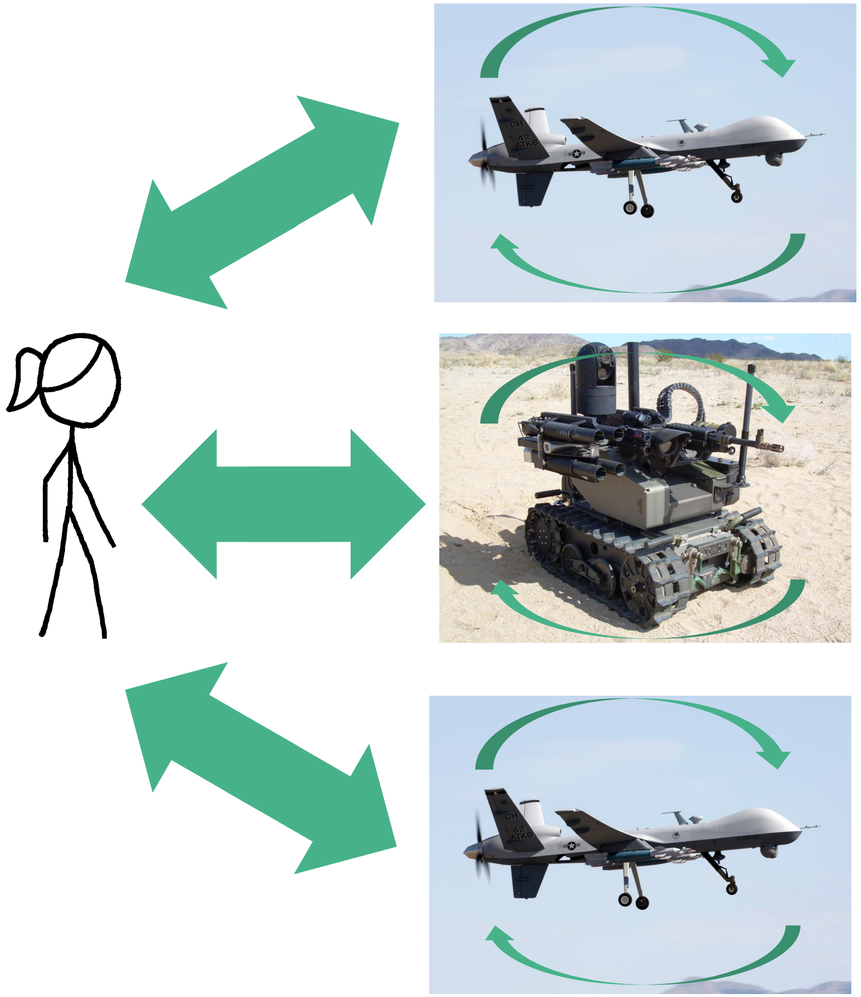
Level 4

Human on top of the loop

FLAVORS OF (MACHINE) GUIDED DECISIONS

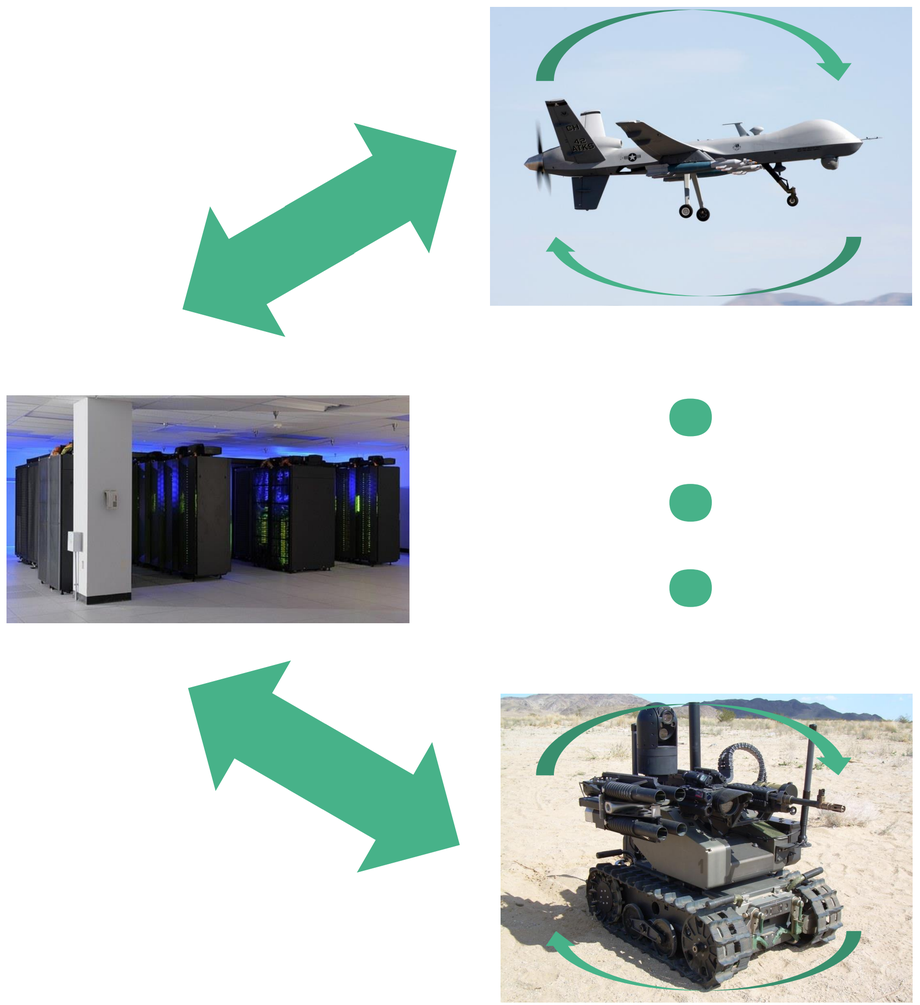
Level 5

Human on top of many loops



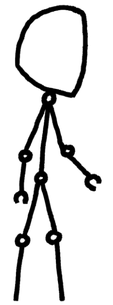
Level 6

AI on top of a great many loops



Level 7

Full machine autonomy



WHY TO ADD HUMANS?/WHAT IS THE HUMAN‘S FUNCTION?

Information & evidence from a broad range of sources

Expert knowledge

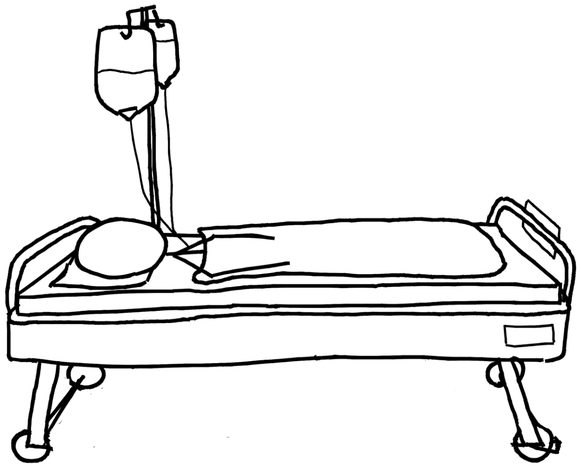
Background knowledge

...

deliberation & judgment

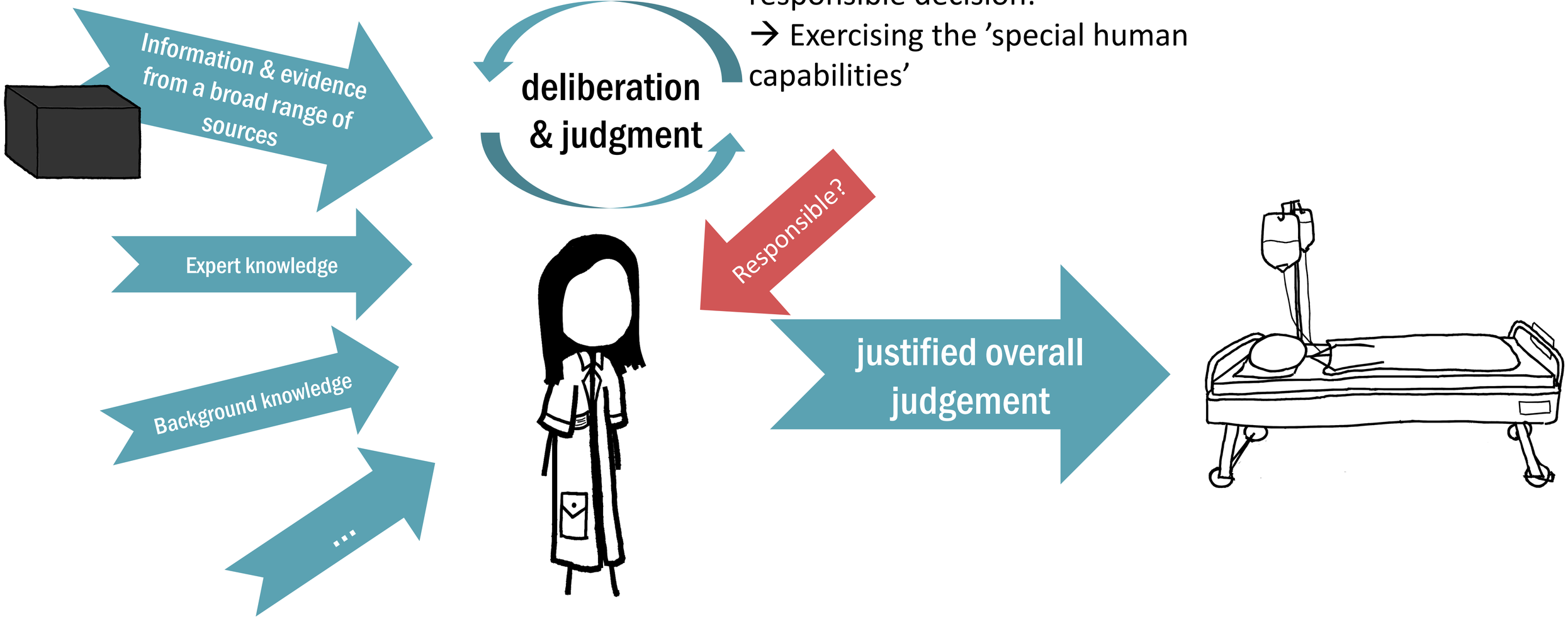
Responsible!

justified overall judgement



WHY TO ADD HUMANS?/WHAT IS THE HUMAN‘S FUNCTION?

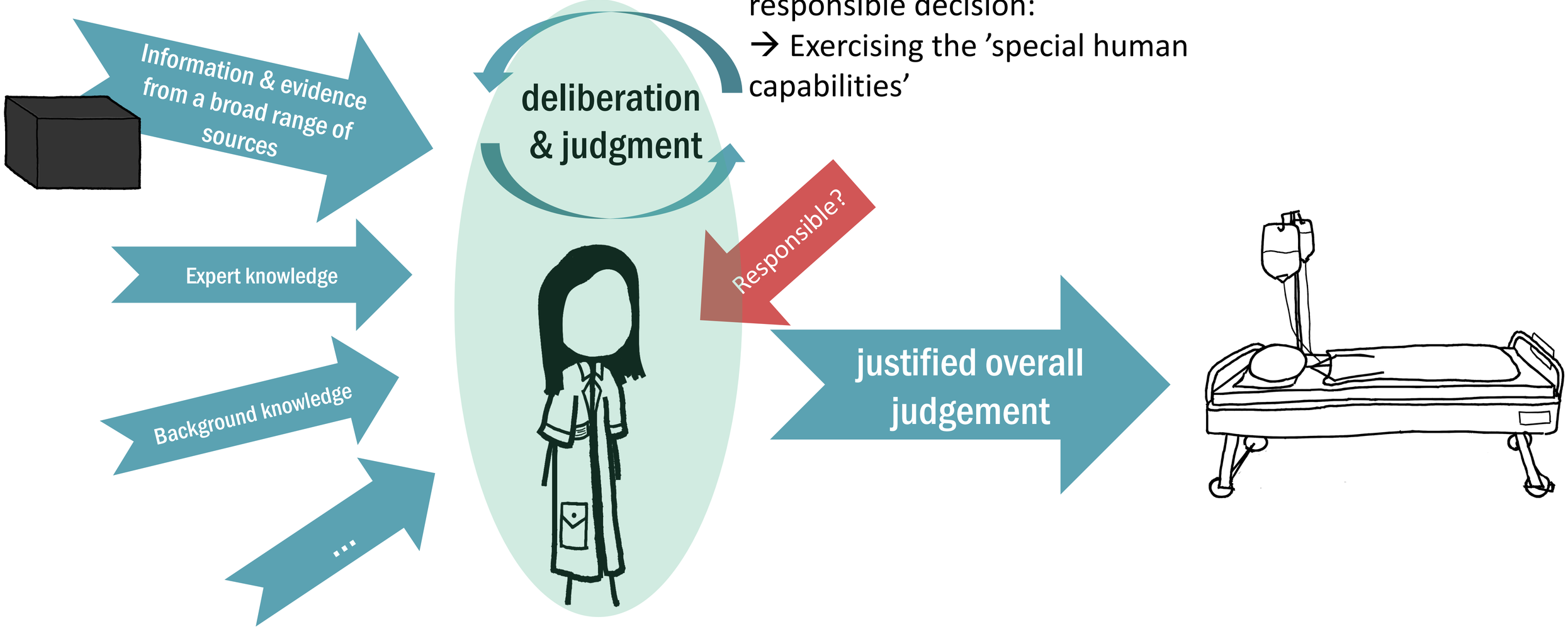
To do what we cannot formalize, quantify or measure and come to a responsible decision:
→ Exercising the ’special human capabilities’



(Under what conditions) bears a human decision-maker
(potentially) moral responsibility
(as being a part in some ASDM)?

WHY TO ADD HUMANS?/WHAT IS THE HUMAN‘S FUNCTION?

To do what we cannot formalize, quantify or measure and come to a responsible decision:
→ Exercising the ’special human capabilities’

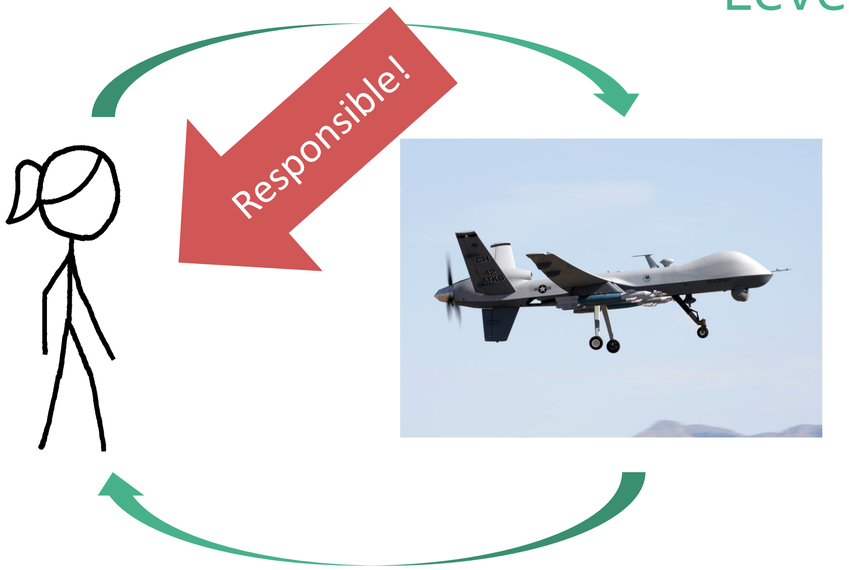


THE NAÏVE APPROACH

"In order to be able to *adequately attribute responsibility*, it suffices that 'the final decision' is made by a human."

Human in the loop

Level 3

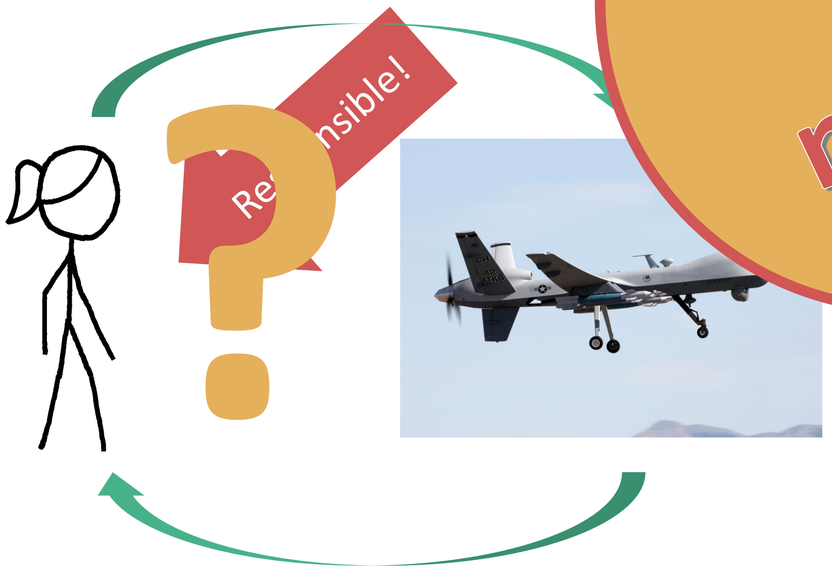


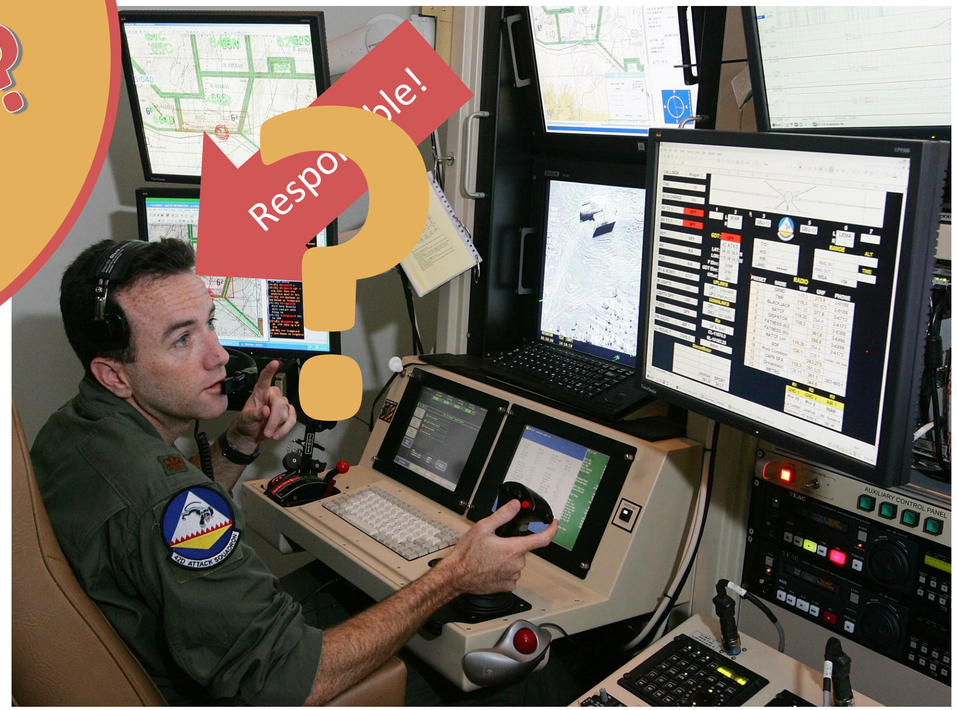
THE NAÏVE APPROACH

"In order to be able to *adequately attribute r*
it suffices that 'the final decision' is mad

too naïve?

Human in the loop





FLAVORS OF (MACHINE) GUIDED DECISIONS

If we only decide and act upon a recommendation like

- “Press the button.”

  or

- “The target is a terrorist with a probability of 86%.”

Are we responsible for these decision and acts?

Or are we just ‘compliant accomplices’?

RECALL: RESPONSIBILITY IN CS

Noorman, Merel, "Computing and Moral Responsibility", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/computing-responsibility/.

For some contexts, they can also be taken as sufficient conditions. For this lecture, we will simplify and say that all three conditions together are sufficient for responsibility.

Prima facie, there are at least three things that are necessary for A being morally responsible for X:

causal connection	epistemic access	control
1 there is a causal connection between A's doing and X	2 A had enough sufficiently well justified beliefs about the possible consequences of their doing that lead to X	3 A had/was able to have some control over their doing that lead to X
Usually, it seems inapt to hold a person responsible for something that is causally independent of her doings.	We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.	If a person could have not done otherwise, it seems inapt to hold her responsible for what she has done.

All of them are potentially problematic in computer science, **but still there are clear cases in which computer scientists and programmers are responsible for their work.**

RECALL: RESPONSIBILITY IN CS

Noorman, Merel, "Computing and Moral Responsibility", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/computing-responsibility/.

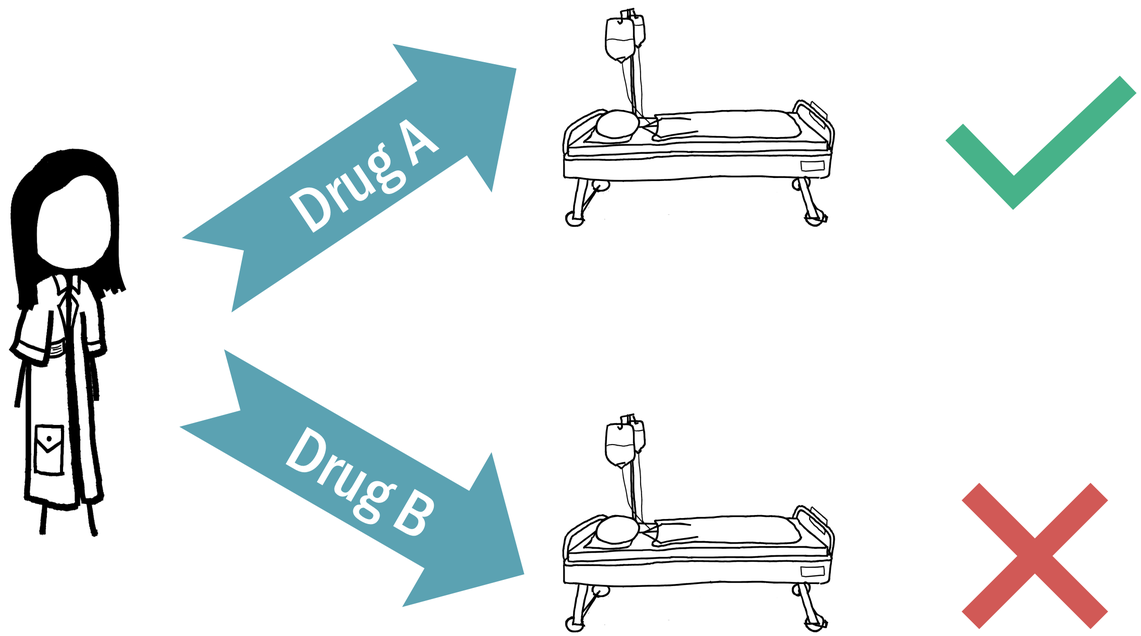
For some contexts, they can also be taken as sufficient conditions. For this lecture, we will simplify and say that all three conditions together are sufficient for responsibility.

Prima facie, there are at least three things that are necessary for A being morally responsible for X:

causal connection	epistemic access	control
1 there is a causal connection between A's doing and X	2 A had enough sufficiently well justified beliefs about the possible consequences of their doing that lead to X	3 A had/was able to have some control over their doing that lead to X
Usually, it seems inapt to hold a person responsible for something that is causally independent of her doings.	We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.	If a person could have not done otherwise, it seems inapt to hold her responsible for what she has done.

All of them are potentially problematic in computer science, **but still there are clear cases in which computer scientists and programmers are responsible for their work.**

WHY EPISTEMIC ACCESS?



epistemic access

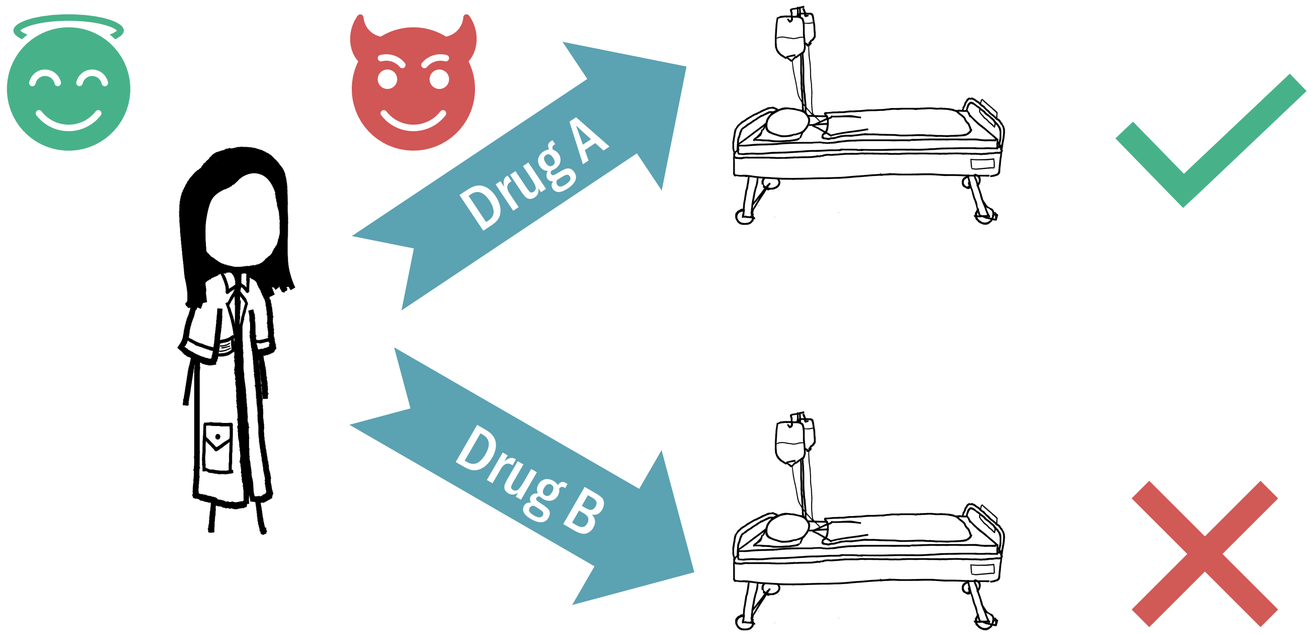
2 A had enough sufficiently well justified beliefs about the possible consequences of their doing that lead to X

We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.

Intuitive appeal

- We want to hold someone responsibility if they did not act in the way they ought to.

WHY EPISTEMIC ACCESS?



epistemic access

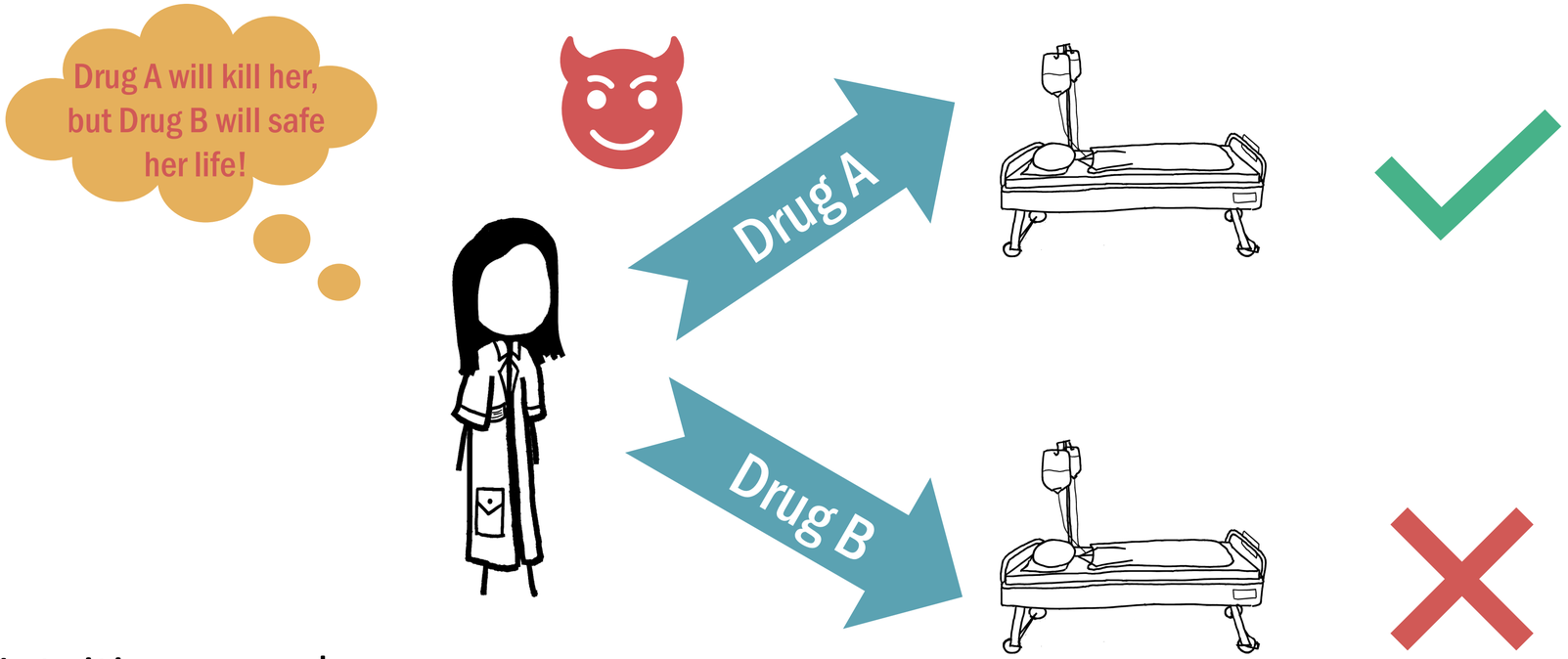
2 A had enough sufficiently well justified beliefs about the possible consequences of their doing that lead to X

We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.

Intuitive appeal

- We want to hold someone responsibility if they did not act in the way they ought to.

WHY EPISTEMIC ACCESS?



epistemic access

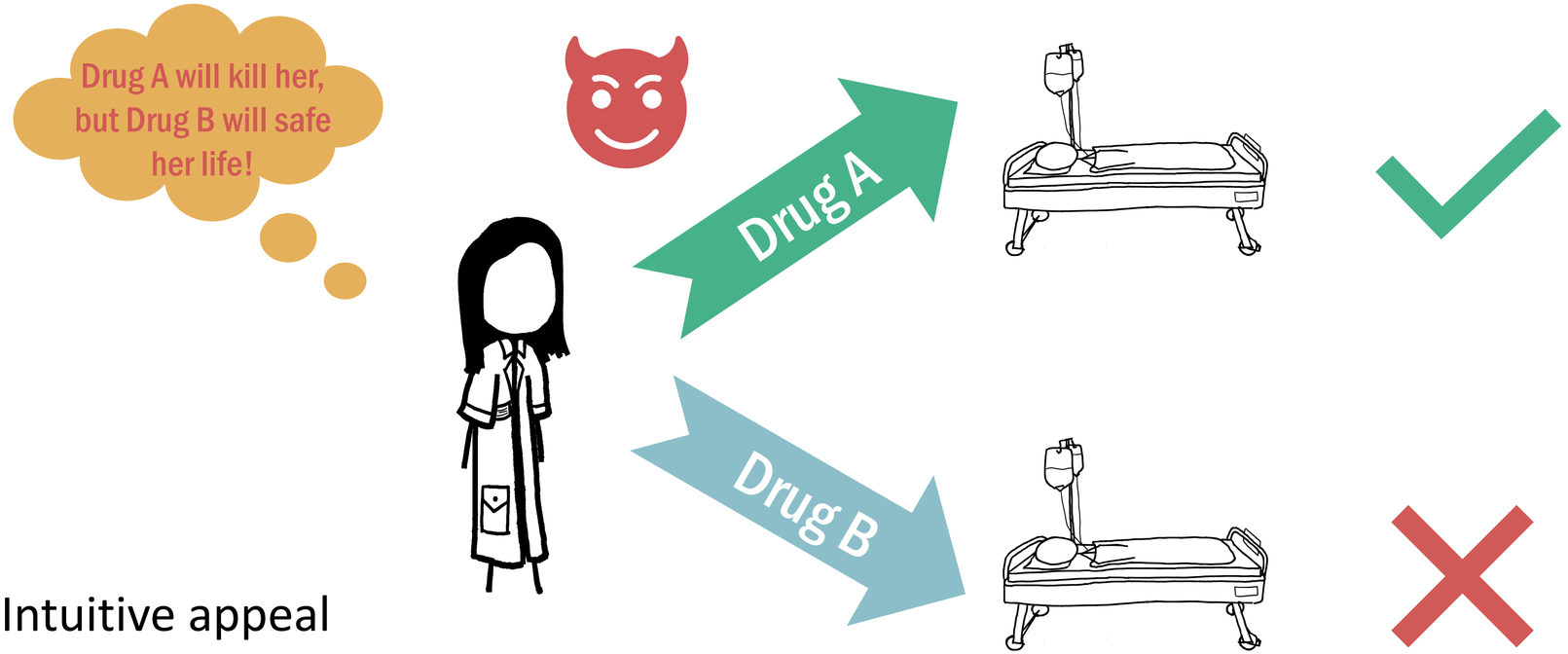
2 A had enough sufficiently well justified beliefs about the possible consequences of their doing that lead to X

We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.

Intuitive appeal

- We want to hold someone responsibility if they did not act in the way they ought to.

WHY EPISTEMIC ACCESS?



epistemic access

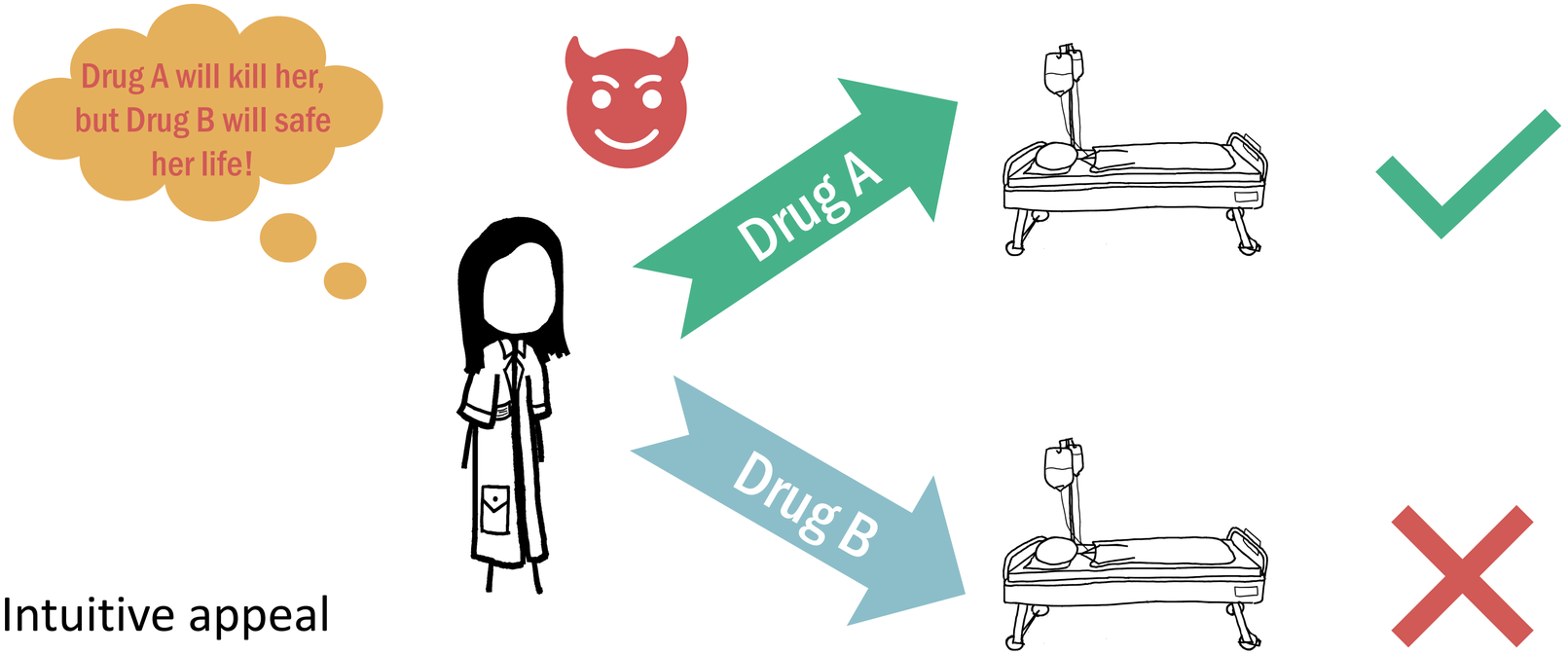
2 A had enough sufficiently well justified beliefs about the possible consequences of their doing that lead to X

We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.

Intuitive appeal

- We want to hold someone responsibility if they did not act in the way they ought to.
- More precisely, we are concerned with the action-guiding (subjective) sense of normativity here:
  - if someone did what they should have believed they ought not to do
    (or if someone did not what they should have believed they ought to do), then they are to blame!

WHY EPISTEMIC ACCESS?



epistemic access

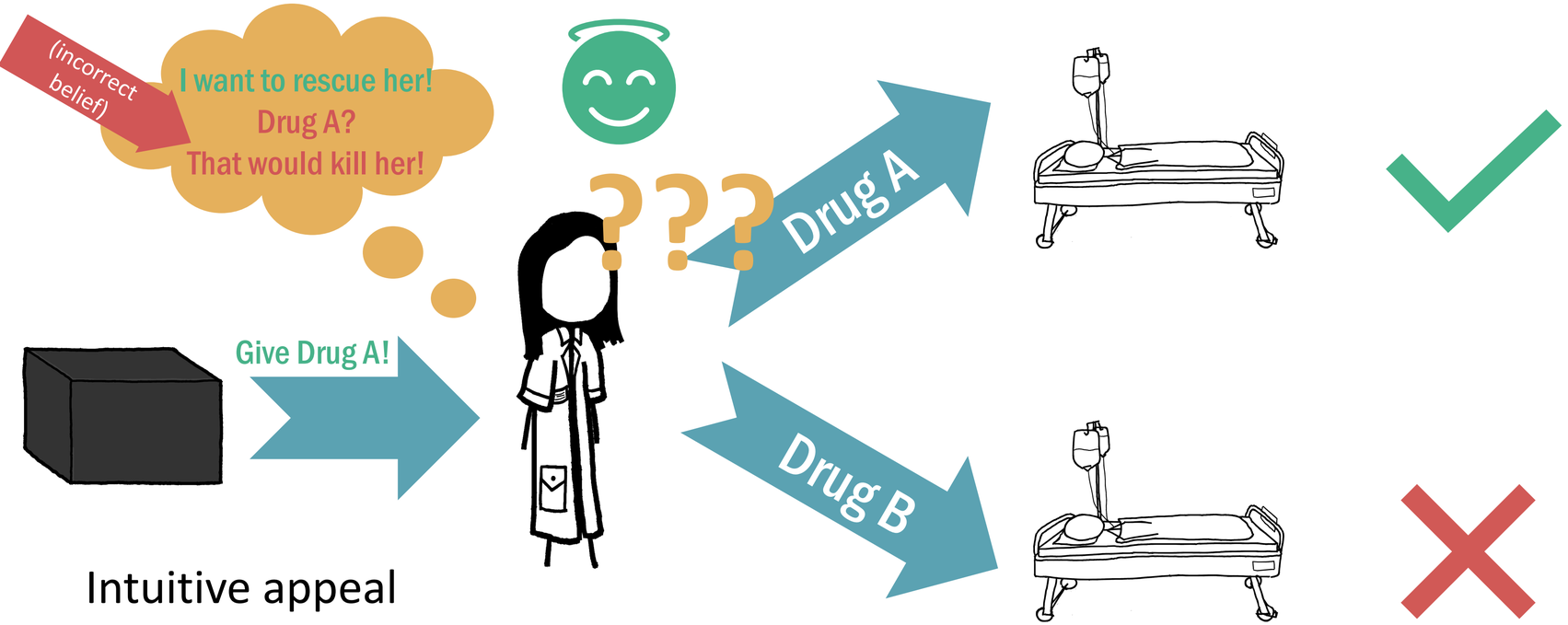
2 A had enough sufficiently well justified beliefs about the possible consequences of their doing that lead to X

We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.

Intuitive appeal

- We want to hold someone responsibility if they did not act in the way they ought to.
- More precisely, we are concerned with the action-guiding (subjective) sense of normativity here:
  - if someone did what they should have believed they ought not to do
    (or if someone did not what they should have believed they ought to do), then they are to blame!
- But without having epistemic access: How can an agent let morals and its principle govern their decision and actions?

WHY EPISTEMIC ACCESS?



epistemic access

2 A had enough sufficiently well justified beliefs about the possible consequences of their doing that lead to X

We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.

Intuitive appeal

- We want to hold someone responsibility if they did not act in the way they ought to.
- More precisely, we are concerned with the action-guiding (subjective) sense of normativity here:
  - if someone did what they should have believed they ought not to do
    (or if someone did not what they should have believed they ought to do), then they are to blame!
- But without having epistemic access: How can an agent let morals and its principle govern their decision and actions?

FLAVORS OF (MACHINE) GUIDED DECISIONS

causal connection

1 there is a causal connection between A's doing and X

✓

epistemic access

2 A had enough sufficiently well justified beliefs about the possible consequences of their doing that lead to X

✗

control

3 A had/was able to have some control over their doing that lead to X

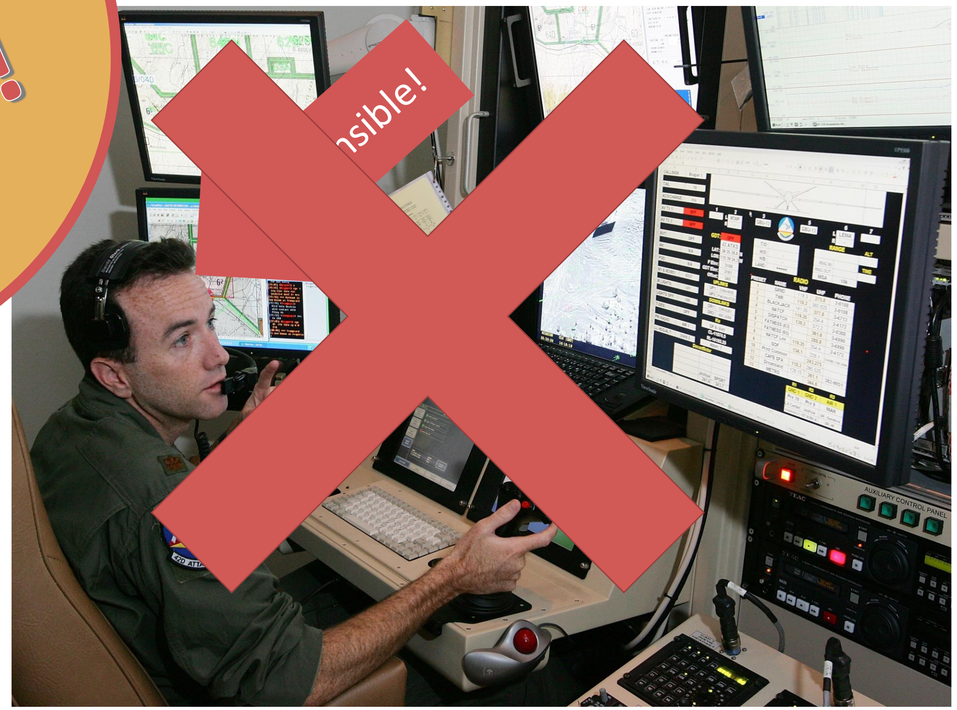
✓

THE NAÏVE APPROACH

"In order to be able to *adequately attribute r*
it suffices that 'the final decision' is mad

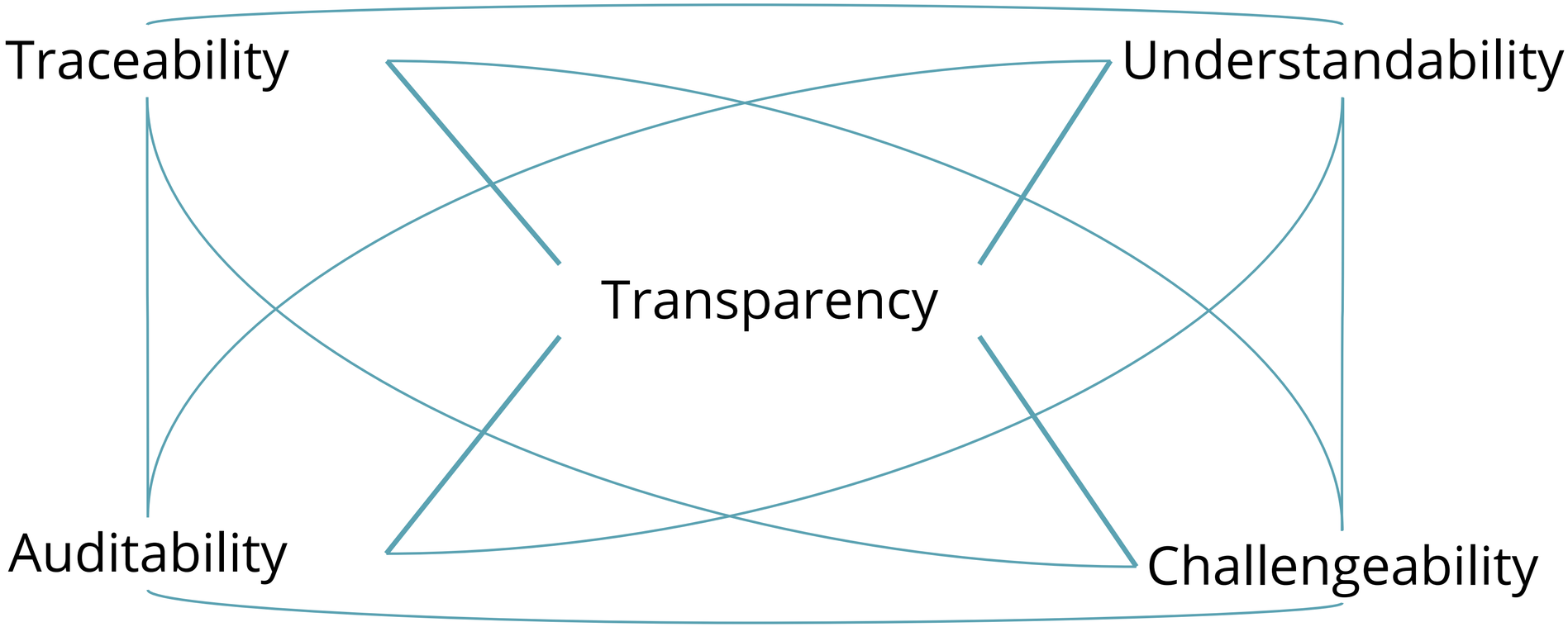
Human in the loop

too naïve!

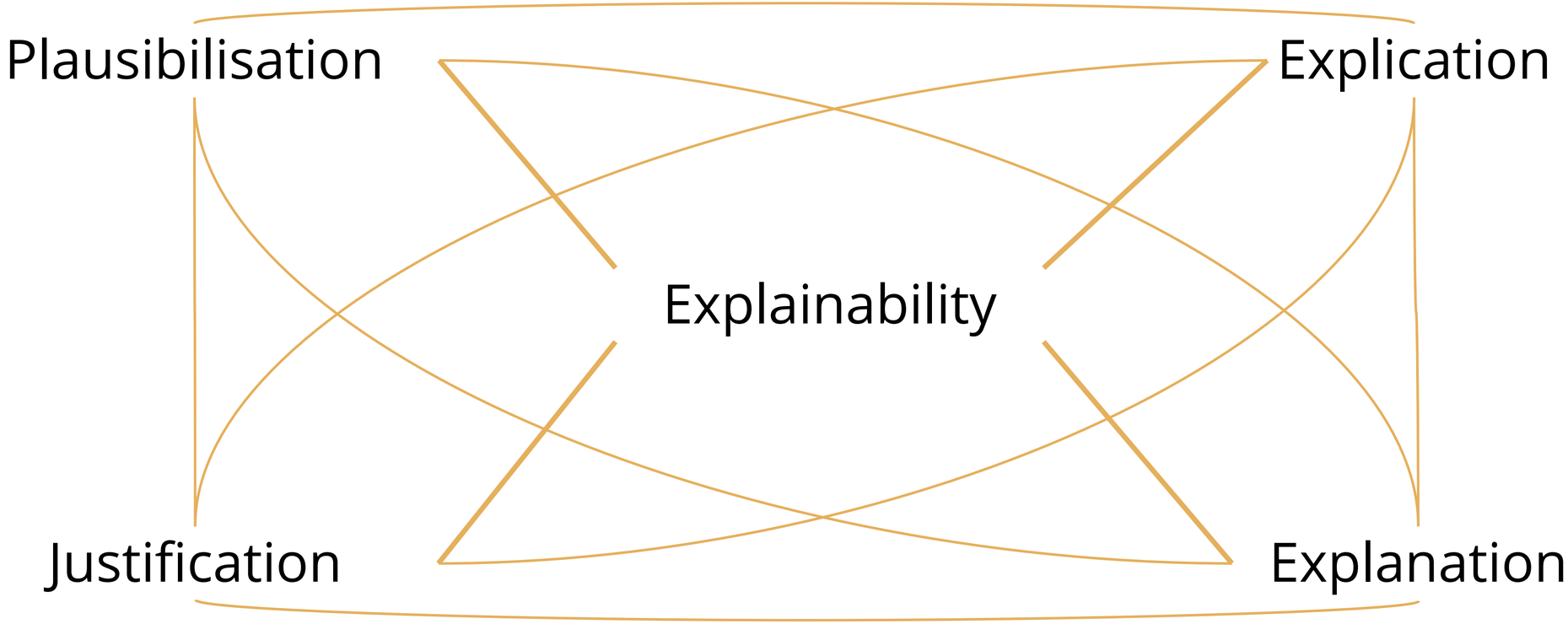


So what?

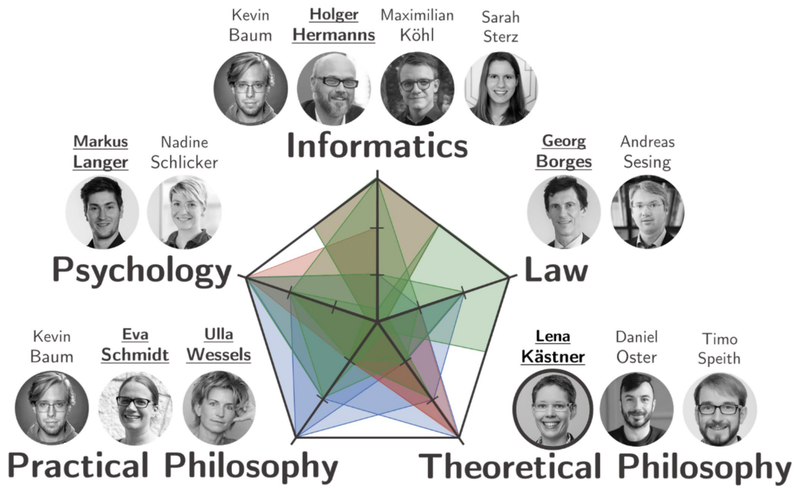
ACCESSIBILITY PROPERTIES



VEHICLES TO ACCESSIBILITY



Interdisciplinary EIS Team



AI in Modern Society

Artificially intelligent systems increasingly augment or take over high-stakes tasks. This trend raises profound moral, legal, and societal challenges. EIS has identified the following societal desiderata:

(i) perceivable trustworthiness,
(ii) competent and responsible decision-making,
(iii) adequate accountability attribution,
(iv) conformity with legal rights, and
(v) preservation of fundamental moral rights.

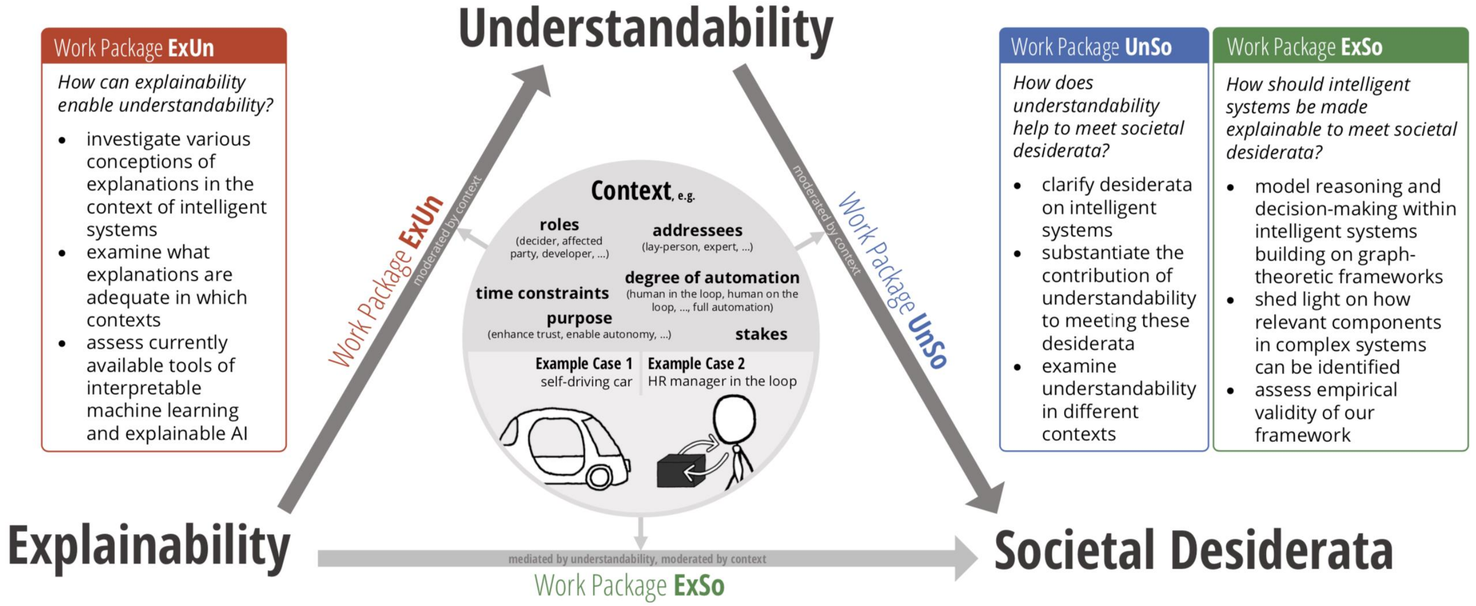
Meeting these desiderata presupposes that specific human stakeholders are able to *understand* intelligent systems' behavior. To enable this understanding, *explainability* is a practical prerequisite.

Research question:
In what ways should intelligent systems and their behaviors be explainable to meet important societal desiderata?

Explainable Intelligent Systems

VolkswagenStiftung

The EIS Approach



heise online › News › 11/2019 › Tödlicher Crash mit autonomem Auto: Fußgänger auf Fahrbahn nicht…

Tödlicher Crash mit autonomem Auto: Fußgänger auf Fahrbahn nicht vorgesehen

In einem Unfall mit einem Roboterauto von Uber kam 2018 eine Frau ums Leben. Das Auto konnte verkehrswidrig die Straße überquerende Personen nicht erkennen.

Lesezeit: 2 Min. In Pocket speichern 1222

Das Unfallfahrzeug (Bild: NTSB)

06.11.2019 10:48 Uhr

Von Martin Holland

https://www.heise.de/newsticker/meldung/Toedlicher-Crash-mit-autonomen-Auto-Fussgaenger-auf-Fahrbahn-nicht-vorgesehen-4578931.html

Center for Perspicuous Computing

Collaborative Research Center 248 Foundations of Perspicuous Software Systems
funded by the DFG

About Us Partners Projects Events Team Contact Career Lecture Series

About

The Transregional Collaborative Research Centre 248 "Foundations of Perspicuous Software Systems" aims at enabling comprehension in a cyber-physical world with the human in the loop.

From autonomous vehicles to Industry 4.0, from smart homes to smart cities – increasingly computer programs participate in actions and decisions that affect humans. However, our understanding of how these applications interact and what is the cause of a specific automated decision is lagging far behind. With the increase in cyber-physical technology impacting our lives, the consequences of this gradual loss in understanding are becoming severe. Systems lack support for making their behaviour plausible to their users. And even for technology experts it is nowadays virtually impossible to provide scientifically well-founded answers to questions about the exact reasons that lead to a particular decision, or about the responsibility for a malfunctioning. The root cause of the problem is that contemporary systems do not have any built-in concepts to explicate their behaviour. They calculate and propagate outcomes of computations, but are not designed to provide explanations. They are not perspicuous.

The key to enable comprehension in a cyber-physical world is a science of perspicuous computing.

CPEC CENTER FOR PERSPICUOUS COMPUTING

perspicuous /pəˈspɪkjʊəs/
adjective *formal*
clearly expressed and easily understood; lucid.
able to give an account or express an idea clearly.

https://www.perspicuous-computing.science/