Ethics and Machine Learning

Machine Learning CSE2510

"AI wondering if it is a tool - DALL-E generated image

Guest lecture - Mark Theunissen, TPM/VTI

Today's Program

1. A Little Taste of Ethics
2. Machine Learning and Ethics

   Break

3. Bias / Fairness
4. Case Study (Lab Assignment)
5. Conclusions / Q&A

Ethics

Who can give an example of an unethical (immoral) behavior or practice?

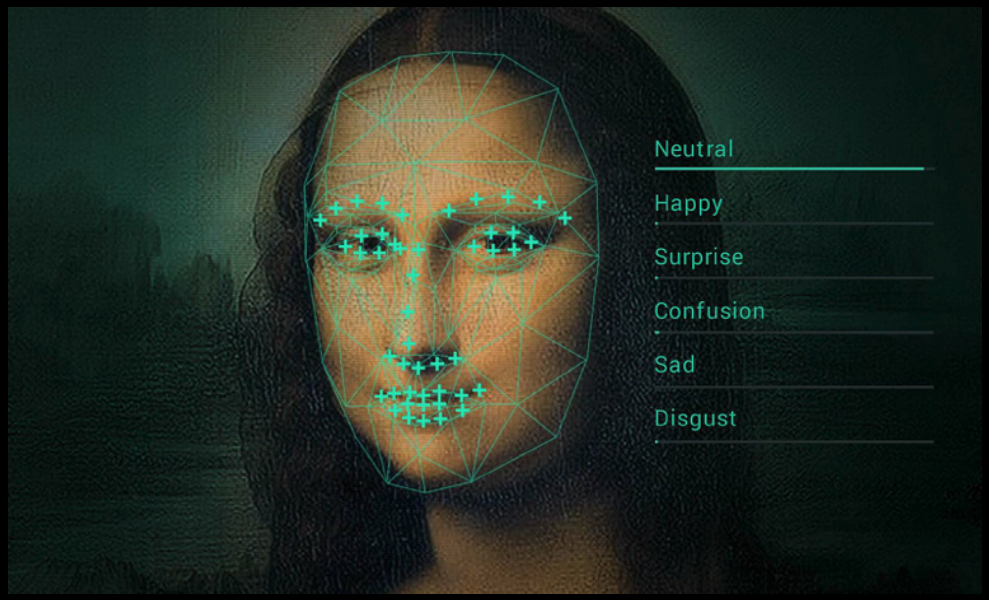
What makes it unethical?

Can you think of an example from your own field, computer science or machine learning?

Humans are Choice Makers

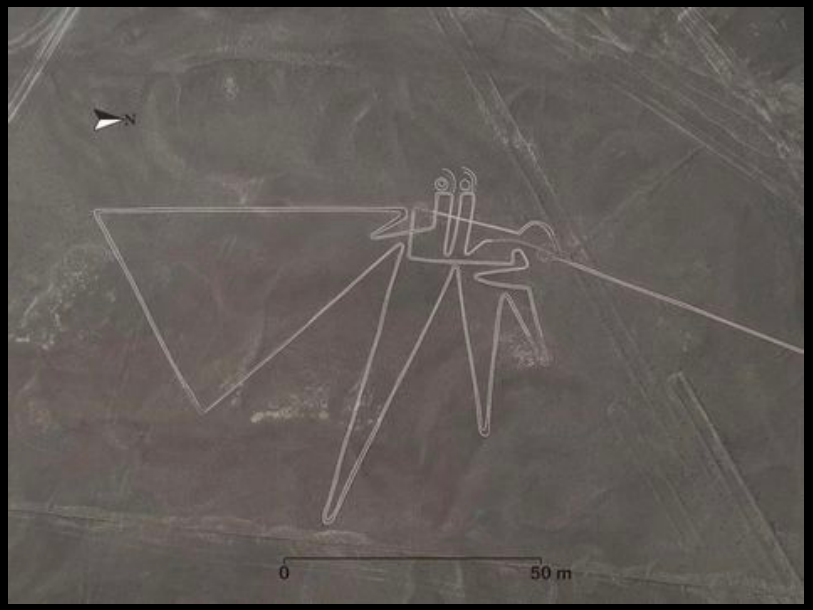
OR

Humans are Choice Makers



Surveillance

OR



Archeology

Ethics

Moral standards of 'right' and 'wrong' that prescribe how we ought to act. Unlike statements of facts that are 'true' or 'false'.

What should I do? / How should I act?

How should I live? What sort of person should I strive to be?

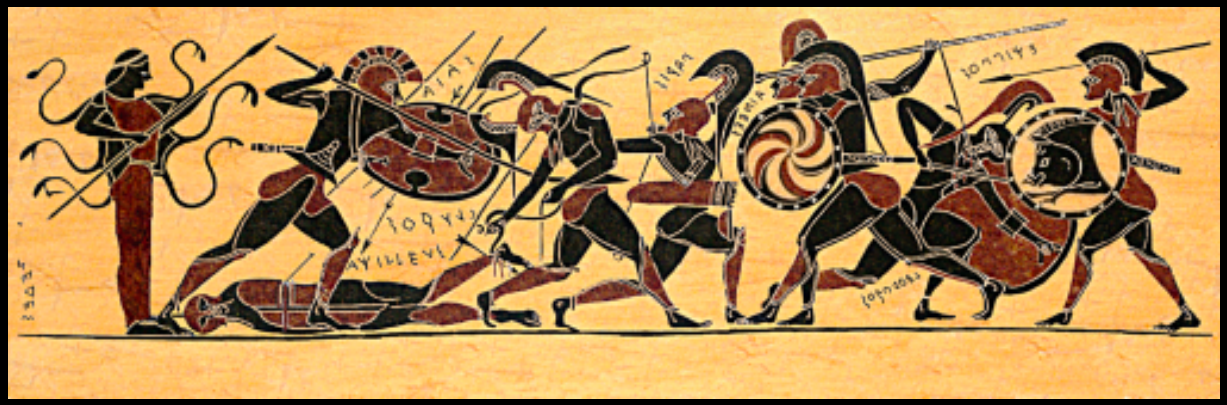
What do we owe to, or how do we treat, those that are affected by our actions and through our practices?

Descriptive / Normative

Descriptive ethics: explanations of, and factual statements about the moral systems people subscribe to, how we act and understand ourselves as moral beings (biology/psychology - anthropology)

Normative ethics: prescriptive accounts on how we ought to act and understand our selves as moral beings. (philosophy)

Descriptive / Normative



‘Home Sapiens is a warring species’

People should not make war: ‘War is Bad!’

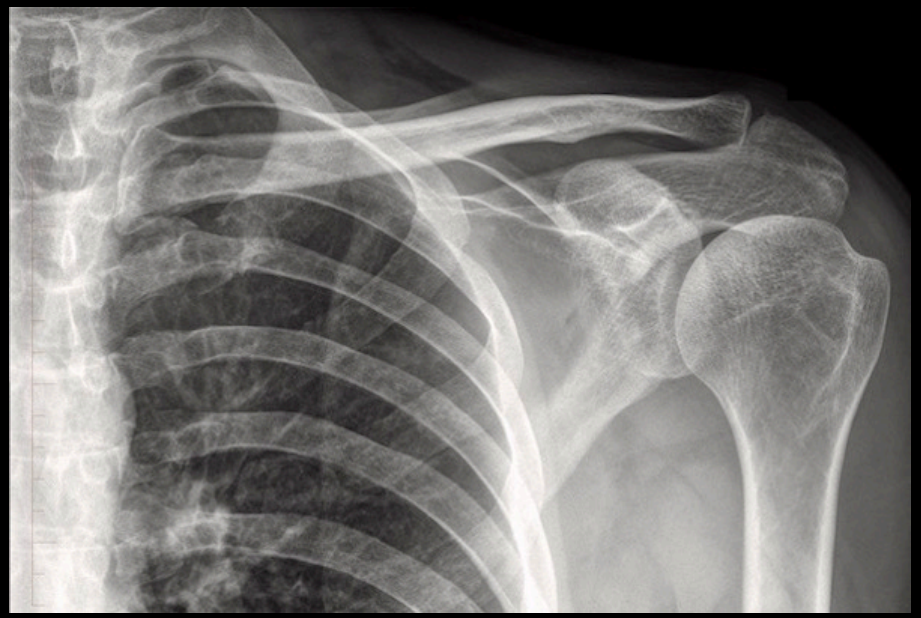
Descriptive / Normative

Note that ML systems can have both 'descriptive' and 'normative' tasks or goals.

Compare:

A: 'Guess someone's age based on the size of their clavicle bone'

B: 'Judge if someone is younger than 18 or not, and should be granted asylum in our country'



Two Common Misconceptions

1. Subjectivism: Moral statement are reflections of merely subjective perceptions and personal preferences

2. Popular opinion: Moral Statements are reflections of public opinion

Subjectivism

Against Moral Subjectivism

We expect others, and others expect us, to give reasons for their ethical choices and beliefs and to argue for why they take those reasons to be correct (to justify the correctness of our belief).

'Two People Arguing About Moral Dilemma Cartoon' - Dall E - 2

Against Moral Subjectivism

Hilary Clinton

AI ‘Gaydar’

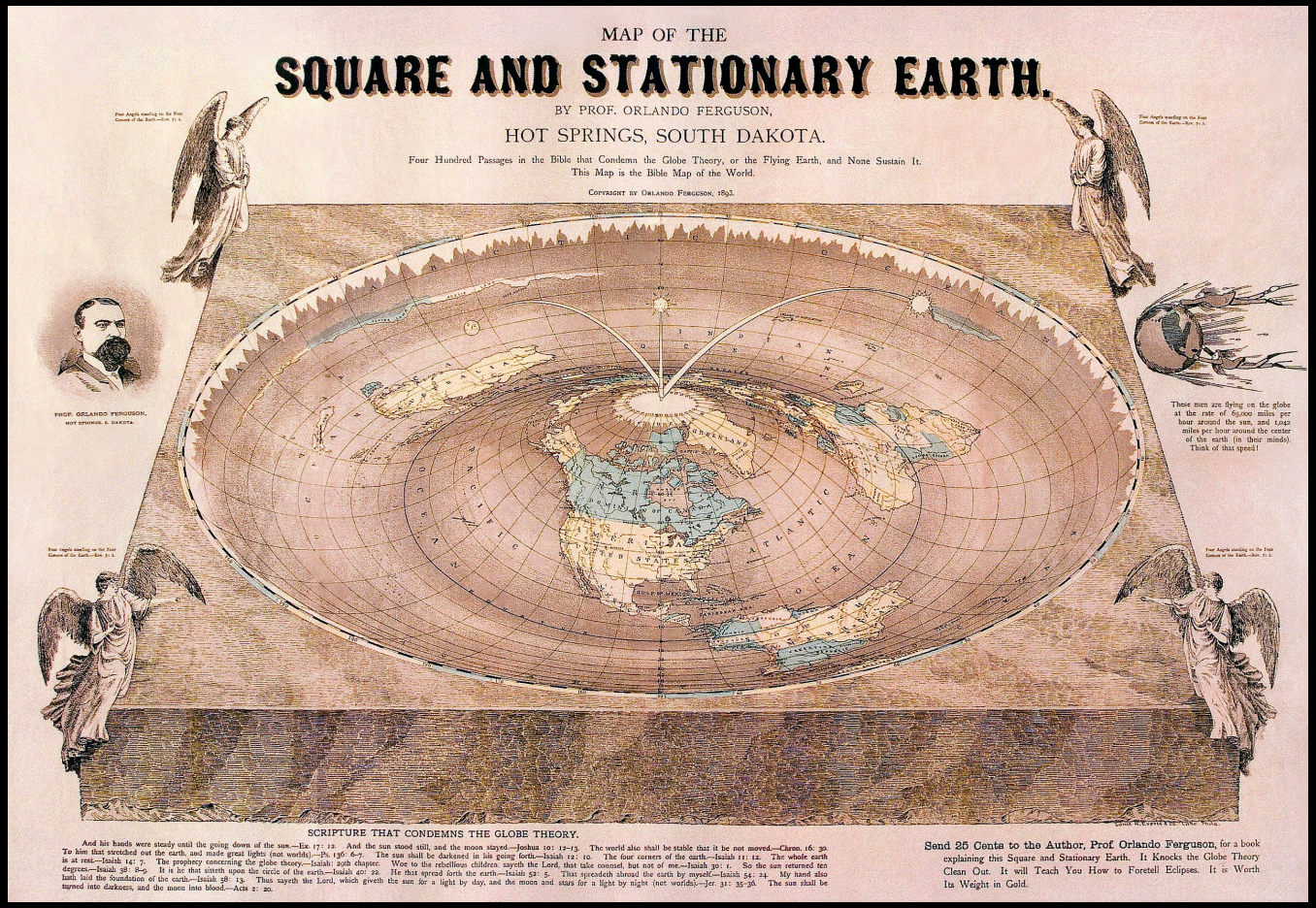
Popular Opinion

2. Popular opinion: Moral Statements are reflections of public opinion

Popular Opinion

Can we answer moral questions by relying on by polling people's opinions or relying on statistics about people's preferences? (the wisdom of the crowd)

Popular Opinion



Against Popular Opinion

Hitler was democratically elected

Jim Crowe (Apartheid) Laws in the United States

Popular Opinion

Some ML developers ask their users if to what extent *the*y think certain (data) features are acceptable to train MLs with

- **Feature 'A priori' fairness: degree to which general population deems the use of a certain features 'fair' without any knowledge about its effect on the outcome**
- **Feature-accuracy fairness: degree to which general population deems the use of a certain feature 'fair' if it increased the accuracy of the outcomes**
- **Feature disparity fairness: degree to which general population deems the use of a certain feature 'fair' despite its use increasing any form of disparity**

Three key ideas in normative ethics

Values > lasting convictions that people believe are desirable to pursue, promote, preserve, or protect. Values are often abstract goals or ideals (e.g., freedom, privacy, well-being, sustainability, fairness, efficiency)

Norms > rules that prescribe which actions are morally required, permitted, or forbidden. Norms that translate values into concrete behaviour (e.g. do not kill, respect others' opinions, do not waste, don't use bad data)

Virtues > character that make someone a good person or that allow people to lead could live (e.g. courage, honesty, generosity, modesty, accuracy, elegance)

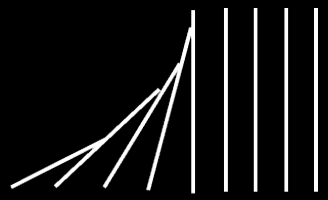
Some Flavors of Normative Ethics

Normative Ethical Theory: Formulation of an account or set of principles that prescribe us how to act

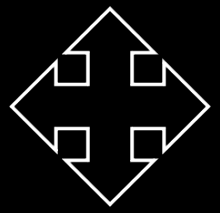
Aims to provide systematic answers to the questions 'How should I live?' and 'How should I act?'

- Consequentialism (Utilitarianism)
- Duty Ethics (Kantian Ethics / Deontology)
- Virtue Ethics

Three Different Points of Emphasis



The consequences of the action (e.g., does it maximize an important value like human happiness?) (If so does the end appears to justify the means)



The action itself and the **Norm/rule** on the basis of which is was chosen (e.g. we must *always*, as a rule, act in such a way that we respect the autonomy in ourselves and others). We act out of duty to the rule, for the sake of the rule that we arrived at through rational procedure.

The agent:
If a person has developed **virtues**, or desirable characteristics (e.g., courage, honesty, empathy), then they can perceive, in each specific context, what the situation demands, ethically speaking. Living ethically is trying to be the best version of yourself.

Consequentialism

- An action is morally required if its consequences maximize happiness and/or minimize suffering
- 'Optimize' to the best outcome (pleasure, happiness, well-being of all)
- Lends itself to quantification and justification of difficult trade-offs between alternatives

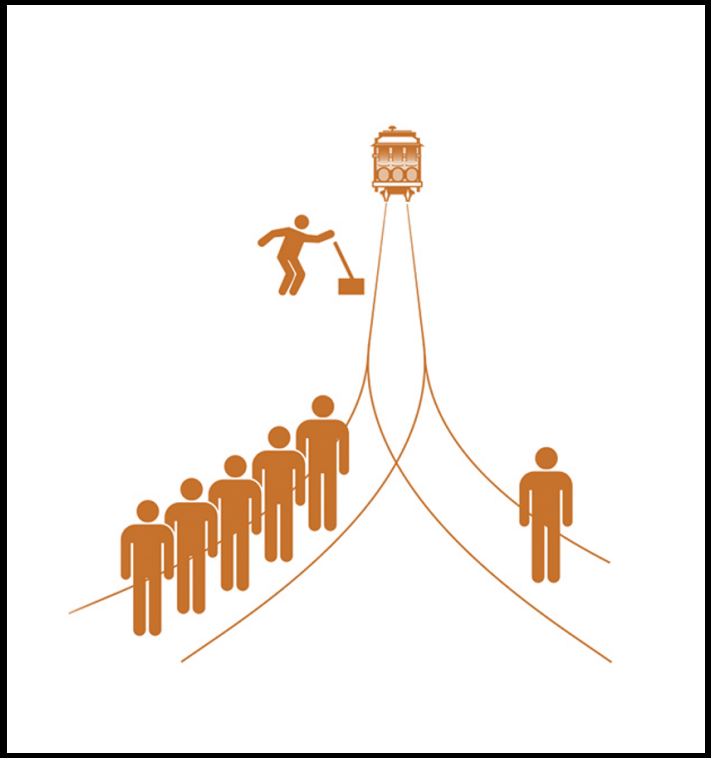
Duty ethics

- You should never use a person as a mere means (a thing, an object, a tool) to some desired end. (Persons are 'end in themselves' and must therefore always be respected)
- Rules for right action *irregardless* of outcomes, some actions are right/wrong because of the kind of action they are
- Often goes with 'rights based' accounts: (e.g.UN declaration of human rights)
- Satisfies our sense of the absoluteness of basic moral principles that won't admit of any exceptions: categorical imperatives

Outcomes vs Basic Rights

These two normative theories are often opposed to each other in thought experiments that test our intuitions about particular moral dilemmas

The Trolley Problem: Is it OK (morally permissible) to pull the lever? Must you pull the lever?



Check the Many Memes!

Can you think of a moral dilemma like this involving a machine learning application? What are the moral concerns at stake?

Note the diversity of domain's of application for ML today and what sort of trade offs are deemed acceptable and why.

Infrastructure/Transport

Law and Criminal Justice

Education

Human Resources

Finance

Fraud Detection

Healthcare

Advertising/Marketing

Natural Sciences

Social Sciences

Taxation

Virtue Ethics

- Previous views are too abstract and general, formulated around one master principle on how to act in each instance, but what would a virtuous do? Can ethics expect actions like these from a person? What sort of person? Do I want to be the kind of person that would and could do that?

- VE focuses on the kind of person we are becoming in the course of the specific lives we live. We learn to be moral over a lifetime of experience, we don't just mechanically adopt a rule, we are habituated to a way of life.

- Virtues of Character: We develop 'good' or 'bad' dispositions to behave under the guide of rational reflection. A good person is someone who…A good doctor is a doctor that…, A good students is a …, A good soldier is a…, a good company is a…, a good ML is a…..

What are virtues of an ML Application?

‘A Noble Artificial Intelligence’ - Dall-E

“A computer trying to find its own blind spots cartoon” - DALL·E

Ethics of Machine Learning

- Ethics of Technology: Goal, Impacts, 'Winners&Losers', Unknowns, Regulation…
- Data Ethics: What is 'good' data (type, content, format)? What (type of) data is prohibited or necessary?
- Algorithmic Ethics: What is a 'just'or 'fair'algorithm? Accuracy, Reliability, Efficiency, Complexity…
- Environmental Ethics: What is a 'sustainable' machine learning system?
- What kind of ML engineers do we want to be…

Carbon FootPrint

The estimated costs of training a model once

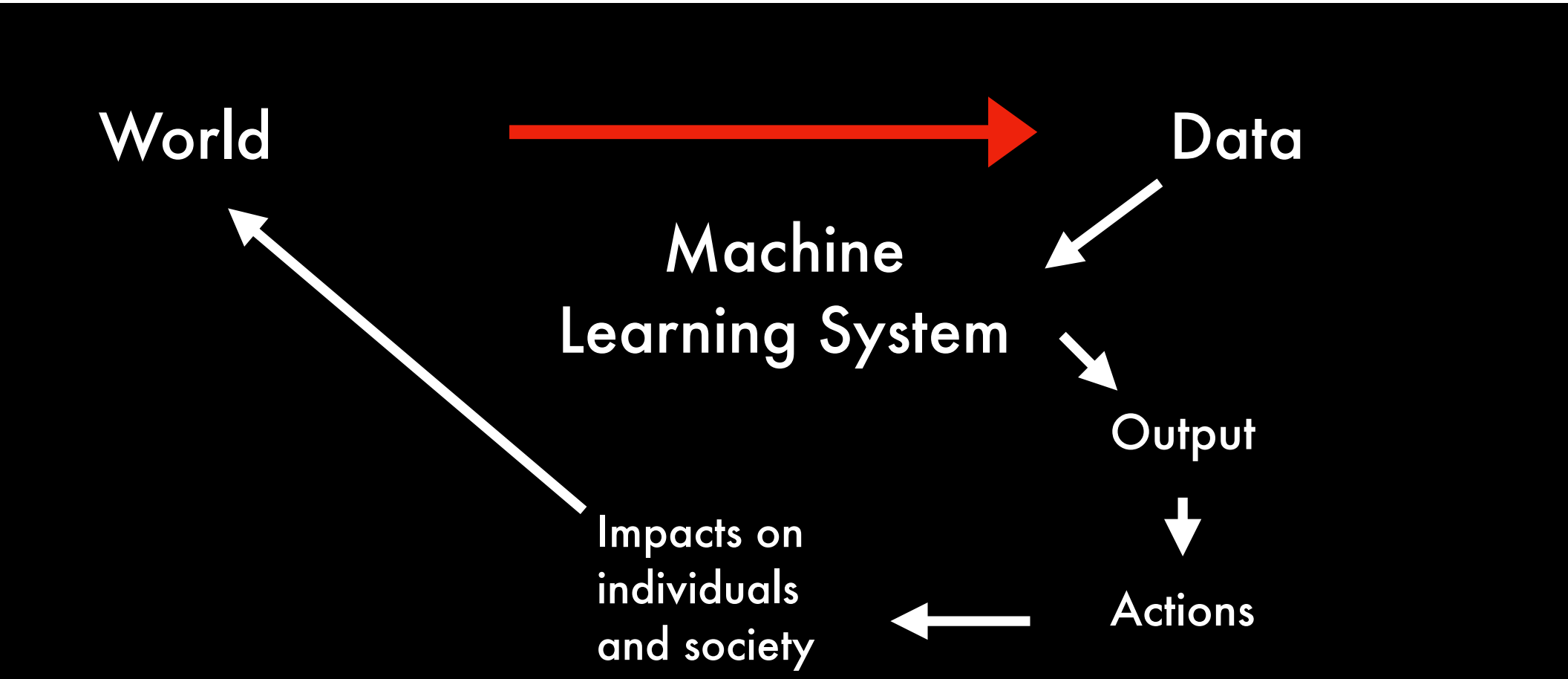
In practice, models are usually trained many times during research and development.

	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud compute cost (USD)
Transformer (65M parameters)	Jun, 2017	27	26	$41-$140
Transformer (213M parameters)	Jun, 2017	201	192	$289-$981
ELMo	Feb, 2018	275	262	$433-$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	$3,751-$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	$942,973-$3,201,722
GPT-2	Feb, 2019	-	-	$12,902-$43,008

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.

Table: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155



ML always exist in a specific and complex social context where specific moral concerns become relevant at different points in the process of building, running, and maintaining the system

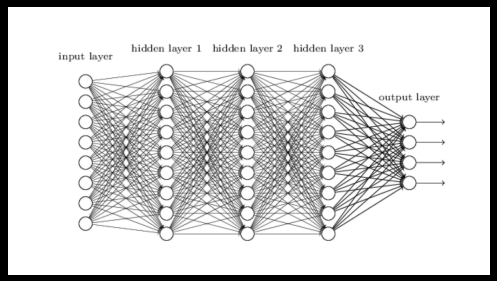
- ✓ Is the goal just?
- ✓ Is the system getting a ‘good training’?
- ✓ Is the system accurate? (note this is extremely complex notion)
- ✓ Is the system reliable?
- ✓ Is the system unbiased?
- ✓ Is the system fair?
- ✓ Is the systems transparent to others, is it explainable?
- ✓ Is the system sustainable?

- ✓ For whom…..? In what context…? In what domain of application…?
- ✓ For how long….?

Injust

Interpretation and Translation (measurement)

Reflects existing injustices, takes in our biases and mistakes



Propagates bias, lacks transparency (black box), features that are proxies for bias, turns on features we think should not be relevant towards outcome

Recommendation or Judgment/Decision: issues of accuracy, reliability, and trust

What actions follow?

What are the actual risks and impacts? For whom? Winners and Losers?

Overall improvement?
Overall harmful?

Where can we intervene to mitigate risks/harms?

What kind of engineers do we want to be…

After the Break

Bias and Fairness

15min Break

Bias

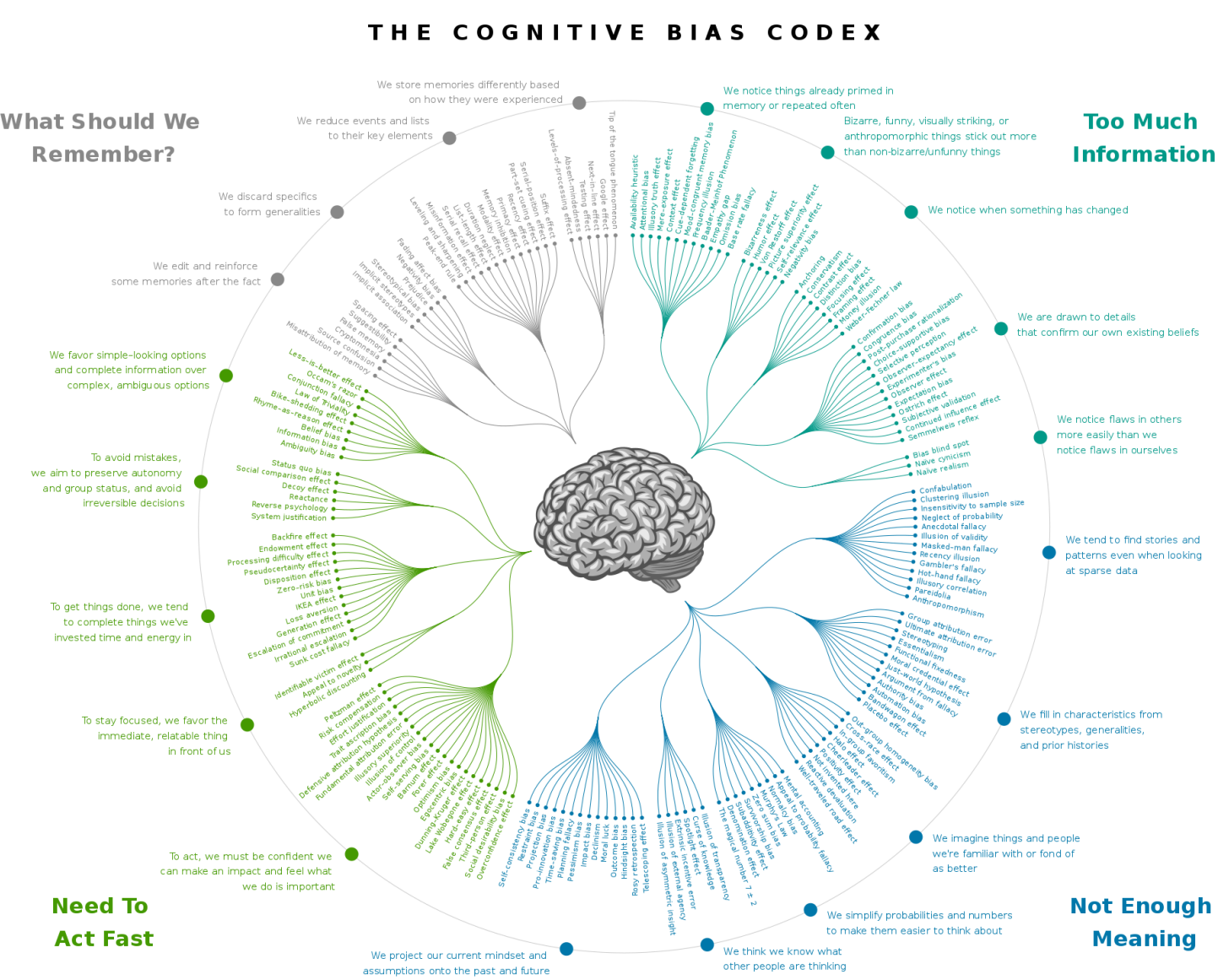
Preference or inclination for or against something. Often accompanied by a tendency to ignore the merits of relevant alternatives or others points of view.

Bias

Can someone share a bias they think they have or exhibit? Is it positive, negative or neutral? Why?

Bias

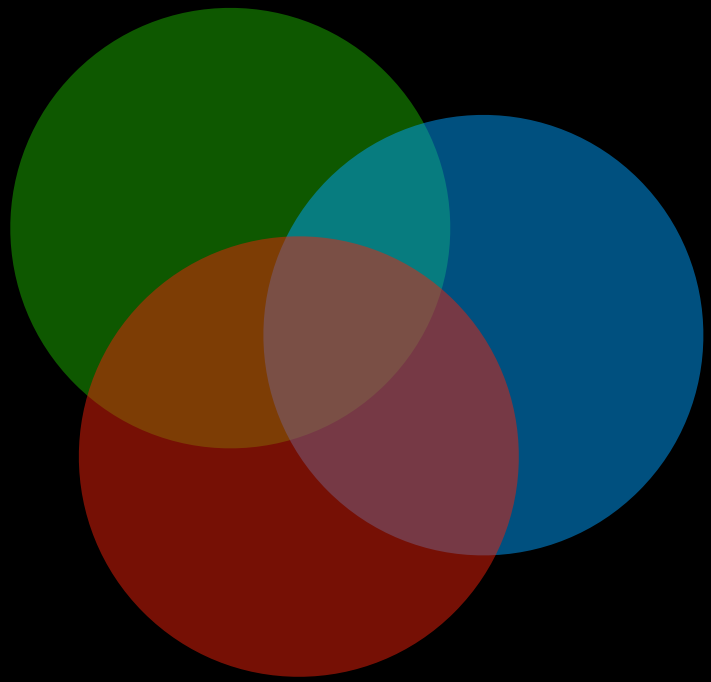
- ✓ They can be ‘neutral’, good or bad. Examples?
- ✓ The can be explicit or implicit. Examples?
- ✓ They can be good or bad (rational/irrational) depending on the context. Examples?
- ✓ They can be good or bad for different reasons. Examples?



There are many types of bias and everyone exhibits some types of bias

Sources of Bias?

- ✓ Natural / Evolutionary
- ✓ Social / Historical
- ✓ Technology / Mediated



When a bias is unreflectively accepted
it can lead us morally astray

Bias → Prejudice → Discrimination

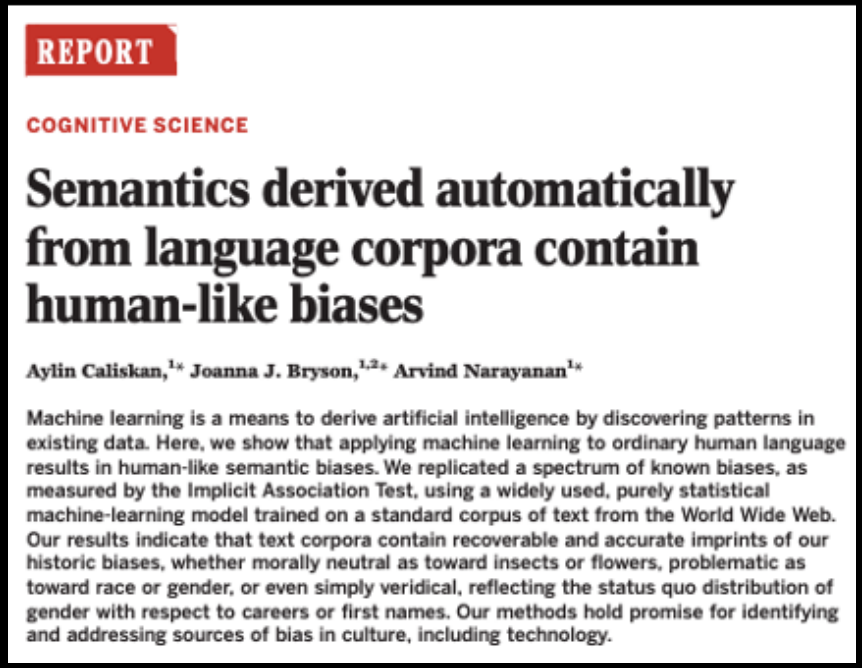
- ✓ Bias, as inclination, can lead to prejudice: (negative) preconceived judgments about others based on inadequate knowledge
- ✓ Prejudice can lead to discrimination: action that treats someone exclusively on the bases of their membership in a category (group) as defined by prejudice

Bias in Machine Learning

- ✓ 'Garbage in Garbage out' - To some degree the world we live in *is* 'garbage' and will be reproduced, made better or worse by applying these kinds of MLs
- ✓ Existing biases can be reproduced and exacerbated by MLs.
- ✓ Algorithms can themselves be biased in ways unbeknownst to us, even if otherwise accurate and reliable
- ✓ The challenge is to mitigate the risk of harmful consequences to a degree 'we' find morally acceptable. To ensure 'fairness.'
- ✓ Who is we? Who is responsible?

Biases can be deeply embedded into our most basic human institutions - like language - and unreflectively reproduced though our technologies

"Our findings suggest that if we build an intelligent system that learns enough about the properties of language to be able to understand and produce it, in the process it will also acquire historical cultural associations, some of which can be objectionable." Caliskan et al., 3



REPORT

COGNITIVE SCIENCE

Semantics derived automatically from language corpora contain human-like biases

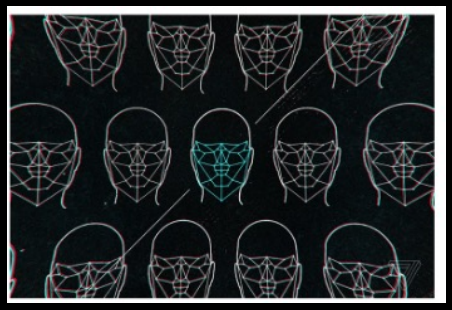
Aylin Caliskan,[1]* Joanna J. Bryson,[1,2]* Arvind Narayanan[1]*

Machine learning is a means to derive artificial intelligence by discovering patterns in existing data. Here, we show that applying machine learning to ordinary human language results in human-like semantic biases. We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names. Our methods hold promise for identifying and addressing sources of bias in culture, including technology.

Bias in Machine Learning

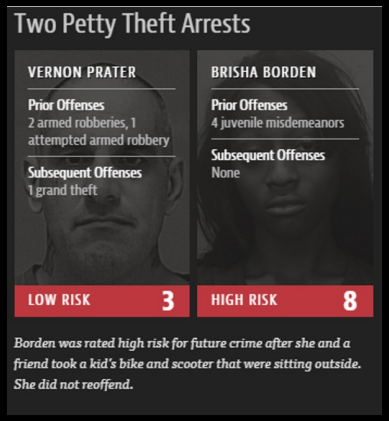
You may have heard of these cases. Can someone remind us of the biases involved here and their source? How could they have been prevented?

Amazon's. Job Recruitement Tool



Amazon's Face recognition Software 'Rekognition'

'Tay' Microsoft Chat Robot



Predictive Policing

Proxy Biases

- ✓ A proxy variable is a variable that is related to a variable of interest to the degree that it can operate as a substitute
- ✓ A proxy bias is then a bias by way of a proxy variable.
- ✓ For example: In the United States for historica reasons, zipcodes are proxies for crime, income, ethnicity etc.
- ✓ Due to the size of data sets and the 'black box' nature of ML systems, proxy biases are likely to occur and sometimes impossible to discover or track before deploying the system in the wild.

Are some of these domains of applications more sensitive than others to a ML application reproducing already existing or news forms of injustice and exclusion?

Law and Criminal Justice

Infrastructure/Transport

Education

Creditworthiness

HR

Healthcare

Targeted Advertising

Fraud Detection

Prediction of Consumer Preferences

(Predictive) Policing

Natural Science

Social Sciences

- Context (domain of application) and occasion (timing and duration of application) matter!

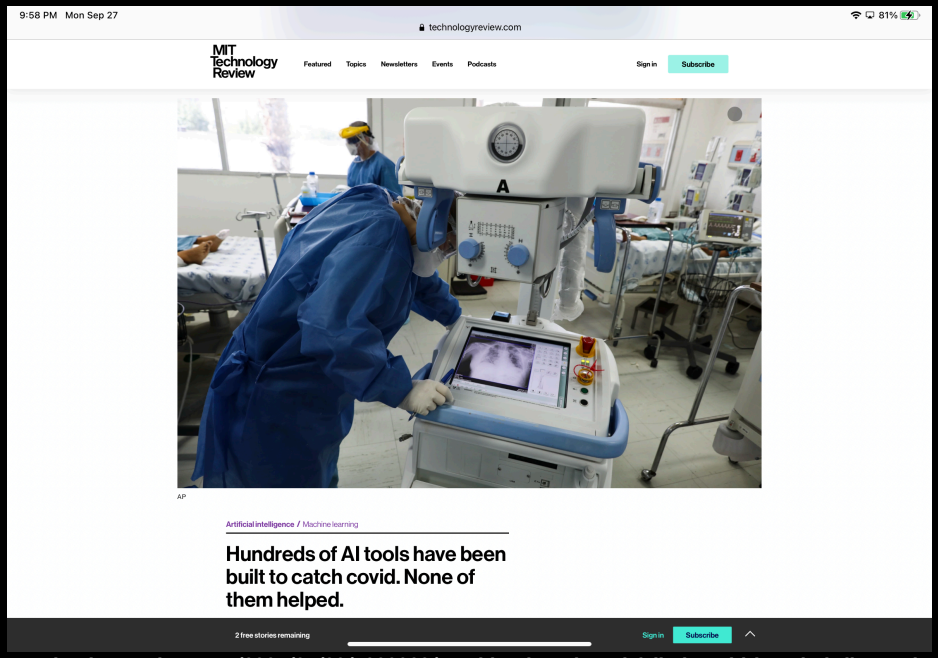
- Shifts in domain and occasion of application can make harmless biases harmful and vice versa!

From Microsoft's statement on Tay:

"For context, Tay was not the first artificial intelligence application we released into the online social world. In China, our XiaoIce chatbot is being used by some 40 million people, delighting with its stories and conversations. The great experience with XiaoIce led us to wonder: Would an AI like this be just as captivating in a radically different cultural environment? Tay – a chatbot created for 18- to 24- year-olds in the U.S. for entertainment purposes – is our first attempt to answer this question."

https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Many ‘Triage ML Applications’ were developed and deployed in the beginning of the Covid19 crises. These helped doctors decide which patients would develop severe symptoms. Given the lack of knowledge about Covid and the urgency to cope with the number of hospitalizations effectively, this appeared justified. However, after several months most of these applications were judged to be more or less useless.



https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

- ML still does not scale well across domains (contexts and occasions) for which it has not been trained
- Biases often ‘appear’ only when an ML is put out in the wild. (Are we then always too late?)
- ML application must always be tested across a representative variety of domains and various occasions/durations to see if it is both accurate and reliable in comparison to other (similar) tools so as to clearly determine its limits
- Question: Consider an ML that we know to be biased against certain groups - say a particular age bias - would it be morally permissible to limit its application to only those groups that are not affected by that particular bias?

Breadth of Sources of Bias in ML

- ✓ Think of all the ways something can go wrong in capturing aspects of the world into data and curating this data into training sets
- ✓ Think of the ways a ML algorithms can 'pick out' certain features as more salient and come to give them more weight (and does this beyond our control?)
- ✓ Think about how a lack of diversity among engineers can affect engineering decisions
- ✓ Think how outside pressures to meet certain goals and to achieve certain results can shape decision making
- ✓ Think how a lack of knowledge of about the causes of certain forms of injustice and bias could effect how ML is built and applied
- ✓ Think how a lack of stakeholder and end-users involvement may affect bias and fairness.

Responsibility and Justification

Who is responsible for preventing or mitigating harmful biases in Machine Learning Applications?

Given the deep social roots of certain biases, are we asking too much from engineers to prevent these? Can they? (Solutionism)

How do we in the end justify the deployment of a ML application in the 'real world'? What sorts of institutions do we need? (FDA for algorithms?)

Conclusion

Justifying the application of ML in the real world is a complex and often morally fraught process that should involve various stakeholders and institutions, adequate testing procedures, law and policy making, specific engineering expertise, variety of user involvement, and a variety of expertise concerning the domains of application.

ML is never ‘a short cut’ solution for a complex problem or decision procedure. MLs have to be integrated into complex systems, practices and institutions, that are themselves often not perfectly just and fair to begin with.

Bias and Fairness

Fairness is about given everyone equal consideration and equal opportunity to benefit in relation to some good or value. (E.g. privacy, free speech, access to services and basic goods)

We want to prevent our biases from undermining our idea of fairness, i.e. to lead to unequal treatment

Different ethical theories have different conceptions of fairness

Fairness and Machine Learning

Fairness in machine learning focusses on ways to prevent the system making biased 'judgments' we consider pernicious; that lack impartiality with respect to a specific vulnerable group or take on features that are otherwise 'morally sensitive', or should be considered altogether irrelevant for a sound judgment.

Particularly hard are proxy biases.

Two Cases: Which is more problematic?

1/ Credit Worthiness: Imagine a bank in the U.S. deploying an algorithm that uses various features, like zipcode, education, income, debts, etc. to determine if you are eligible for a loan.

2/ Medical Diagnoses: Imagine an AI that determines your risk of diabetes using features like age, sex, race, lifestyle, diet

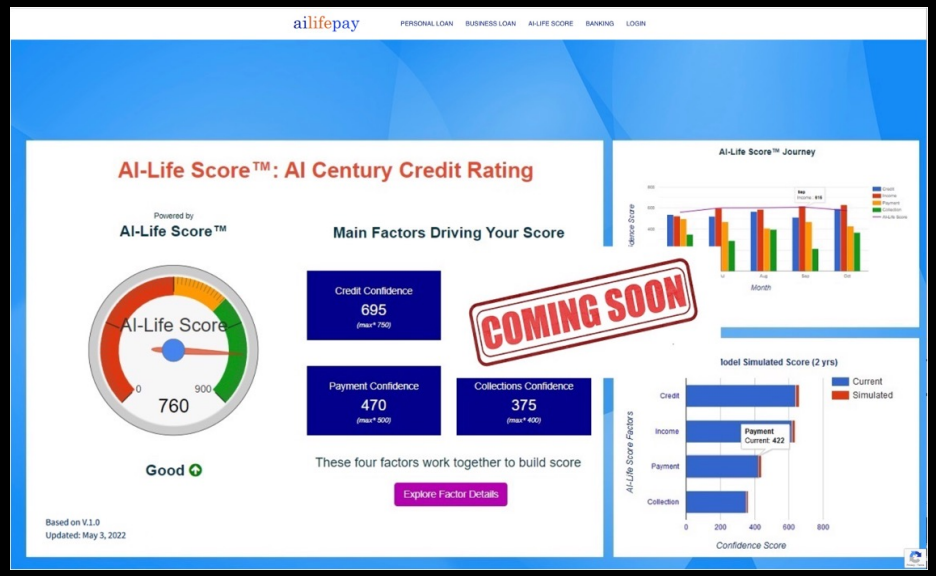
Fairness in the training data?

- Outcome Fairness: The outcome of the system should be fair.
- Process Fairness: The process through which an outcome is reached should be fair

Fairness in training data?

- Fairness through 'unawareness': simply don't us protected feature
- Fairness as statistical parity: prediction is independent from specific protected feature
- Feature 'A priori' fairness: degree to which general population deems the use of a certain feature 'fair' without any knowledge about its effect on the outcome
- Feature-accuracy fairness: degree to which general population deems the use of a certain feature 'fair' if it increased the accuracy of the outcomes
- Feature disparity fairness: degree to which general population deems the use of a certain feature 'fair' despite its use increasing any form of disparity

Lab Assignment



AI Credit Rating Tool – 'AI life pay'- https://ailifepay.com/

In the lab assignment you will have an opportunity to practice with machine learning fairness on a small dataset describing bank loans. Here, the task is to predict whether a person will be able to repay their loan; this information will be used to decide whether they should receive money from the bank or not.

Lab Assignment

Four cases that 'model' different notions of fairness as statical parity.

Independence :

1. Ensure that each group is treated by the same criteria and equally represented in the outcomes regardless of 'accuracy'
2. Ensure that each group is equally represented in the outcome but judged by different criteria if that results in better 'accuracy'

Separation Criteria:

3. Fraction of true positives for each group should be the same
4. Fraction of false positives for each group should be the same

The latter may require adjusting the performance of the algorithm over one group and not the other, and thus turning explicitly on how a protected feature is processed

Is this group level approach fair?

Questions?

Contact me: m.r.theunissen@tudelft.nl