SBMT5710
HKUST – Spring 2020

Introduction to Ethics & Fairness in AI and Big Data

https://home.cse.ust.hk/~golin/Talks/mba_golin_spring_2020_AI_Ethics.pdf
https://home.cse.ust.hk/~golin/Talks/Ethics_in_AI_Reading_List.html

Mordecai Golin
Dept. of Computer Science & Engineering
Hong Kong UST

Version 3.2

A quick-and-dirty informal introduction to ethical and fairness issues that arise with the use of AI and big data processing into the commercial workplace.

Some obvious. Some surprising

Also, a few implications for legal obligations (anti-discrimination legislation) and avoiding public relations disasters.

Quick and Dirty because

- Ethics will not be formally defined.
- Very little math. Mostly anecdotal.

See additional reading list for more detailed follow-ups

Some Topics
From Speculative to the Here & Now

- Robot Rights
  - Humane treatment of AIs?
- Unemployment & Inequality Arising from Introduction of AI
  - How do we deal with people losing jobs
- The Ethics of Autonomous Weapons
  - *Humans in the loop* vs *Humans on the loop*
- The Ethics of Experimenting on Humans
  - Facebook's emotion experiment
- Autonomous Vehicles (self driving cars)
  - The ethics that need to be programmed into the system
- Fairness in Machine Learning
  - ML easily reinforces biases

Robot Rights

- Should “intelligent machines” have civil rights?
- Not there yet, but it’s quite possible that in foreseeable future we will build software that will at least mimic sentience
  (If not at human level, at some animal level)
- “If you’ve got a computer or a robot that’s autonomous and self-aware, I think it would be very hard to say it's not a person,”
  Kristin Andrews *York University Research Chair in Animal Minds.*
- Something to think about….

I did say we’re starting with the speculative….

Some Topics
From Speculative to the Here & Now

- Robot Rights
  - Humane treatment of AIs?
- Unemployment & Inequality Arising from Introduction of AI
  - How do we deal with people losing jobs
- The Ethics of Autonomous Weapons
  - *Humans in the loop* vs *Humans on the loop*
- The Ethics of Experimenting on Humans
  - Facebook's emotion experiment
- Autonomous Vehicles (self driving cars)
  - The ethics that need to be programmed into the system
- Fairness in Machine Learning
  - ML easily reinforces biases

Unemployment & Inequality Arising from Introduction of AI

Anton Korinek, Joseph E. Stiglitz
"Artificial Intelligence and Its Implications for Income Distribution and Unemployment"
NBER Working Paper No. 24174, December 2017

- Stiglitz emphasizes distinction between
  - AI that helps people do their job better
  - AI that replaces/displaces workers

- AI accelerating century old trend of automation displacing jobs
- What does society "owe" its citizens in terms of employment

….. we provide several simple economic models to describe how policy can counter these effects, even in the case of a "singularity" where machines come to dominate human labor. Under plausible conditions, non-distortionary taxation can be levied to compensate those who otherwise might lose.

Unemployment & Inequality Arising from Introduction of AI

- AI accelerating century old trend of automation displacing jobs
- What does society "owe" its citizens in terms of employment

….. we provide several simple economic models to describe how policy can counter these effects, even in the case of a "singularity" where machines come to dominate human labor. Under plausible conditions, non-distortionary taxation can be levied to compensate those who otherwise might lose.

Can think of this as extension of the ***Crisis of Capitalism***,

or ***Socialism***, or whatever your favorite economic model is….

Way beyond the remit of this course but this IS an ethical issue arising from the use of AI that will need to be addressed. Soon.

Some Topics
From Speculative to the Here & Now

- Robot Rights
  - Humane treatment of AIs?
- Unemployment & Inequality Arising from Introduction of AI
  - How do we deal with people losing jobs
- The Ethics of Autonomous Weapons
  - *Humans in the loop* vs *Humans on the loop*
- The Ethics of Experimenting on Humans
  - Facebook's emotion experiment
- Autonomous Vehicles (self driving cars)
  - The ethics that need to be programmed into the system
- Fairness in Machine Learning
  - ML easily reinforces biases

The Ethics of Autonomous Weapons

- *Lethal Autonomous Weapons (LAWs)*
- Three Different Types
  a) Humans In the Loop (HITL)
     Humans needed to make kill decisions
  b) Humans on the Loop (HOTL)
     Humans can step in and abort decisions
  c) Humans Out of the Loop
     Exactly what it says

Is constructing an HOTL LAW ethical?

The Ethics of Autonomous Weapons

- *Lethal Autonomous Weapons (LAWs)*

  Is constructing an HOTL LAW ethical?

- UN recently (2018) started trying to create rules governing LAWs as part of its Convention on Conventional Weapons (CCW)
- Russia is boycotting initiative, so movement on this is unlikely

- LAWs probably already exist.
- Samsung SGR-A1 sentry gun patrols border with North Korea
  - Serious studies have concluded that the SGR-A1 is HOTL capable.

The Ethics of Autonomous Weapons (update)

- Slides on previous page were from last year
- The world moves very fast
  - In a March 2019 UN meeting, Russia was joined by the US, Australia, the UK and Israel in opposing a Law ban
  - Meanwhile, a Chinese company (Ziyan) is exporting a (helicopter) drone advertised "*as capable of full autonomy, including the ability to conduct lethal targeted strikes*."

The Ethics of Autonomous Weapons (More)

- At 1st (and 2nd) glance HOTL LAWS seem like a very bad idea.
- There is a small contingent of ethicists who argue the opposite.

Ronald Arkin. *The case for banning killer robots: counterpoint*. CACM, Dec. 2015

> *"I have the utmost respect for our young men and women in the battlespace, but they are placed into situations where no human has ever been designed to function…..*
> *It is not my belief that an unmanned system will ever be able to be perfectly ethical in the battlefield, but I am convinced they can ultimately perform more ethically than human soldiers."*

The Ethics of Autonomous Weapons (More)

- At 1st (and 2nd) glance HOTL LAWS seem like a very bad idea.
- There is a small contingent of ethicists who argue the opposite.

S. Umbrello, P. Torres and A. F. De Bellis
The future of war: could lethal autonomous weapons make conflict more ethical?
AI & SOCIETY (2020) 35:273–282

- *"…. we contend that the relatively low cost of LAWs, their potential for moral programming, and their ability to remove human combatants from the line of fire constitute strong reasons for pursuing the development and use of LAWs in conflict situations."*

- *"Even more, we argue that "moral LAWs" could constitute the only entities capable of making genuinely ethical decisions about whether its targets live or die."*

Their Explicit Caveats:
- LAWs must have targeting and judgment systems that are equal or superior to the targeting abilities of humans
- The LAWS must embody a moral program or programs that all parties agree upon.

Some Topics
From Speculative to the Here & Now

- Robot Rights
  - Humane treatment of AIs?
- Unemployment & Inequality Arising from Introduction of AI
  - How do we deal with people losing jobs
- The Ethics of Autonomous Weapons
  - *Humans in the loop* vs *Humans on the loop*
- The Ethics of Experimenting on Humans
  - Facebook's emotion experiment
- Autonomous Vehicles (self driving cars)
  - The ethics that need to be programmed into the system
- Fairness in Machine Learning
  - ML easily reinforces biases

Ethical Human Experimentation in Big Data

- Social Scientists (usually) hew to a rigorous set of ethical standards when running experiments on humans
- These include *informed consent* and *allowing opt-outs*
- Oversight is usually provided by university research offices and granting agencies

Would it be ethical for Facebook to experiment on half a million people without getting informed consent?

They did!

The Facebook Mood Experiment

- Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock
  "Experimental evidence of massive-scale emotional contagion through social networks"
  Proceedings of the National Academy of Sciences of the United States of America (PNAS)
  June 17, 2014 111 (24) 8788-8790

- Robinson Meyer
  "Everything We Know About Facebook's Secret Mood Manipulation Experiment"
  The Atlantic
  June 28, 2014

The Facebook Mood Experiment

``For one week in January 2012, data scientists skewed what almost 700,000 Facebook users saw when they logged into its service.

Some people were shown content with a preponderance of happy and positive words; some were shown content analyzed as sadder than average.

And when the week was over, these manipulated users were more likely to post either especially positive or negative words themselves.''

Note: Prior to this there had been research on *correlation*.
For example, did people who saw happy posts use positive words in their own posts?
This was different. This was experimentation that involved manipulating people!

The Facebook Mood Experiment

Note: based on Facebook's terms of use, this was probably legal.
The question is whether it was ethical.

PNA's own editorial statement of concern summarized the issues very well

Questions have been raised about the principles of informed consent and opportunity to opt out in connection with the research in this paper. The authors noted in their paper, "[The work] was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research." When the authors prepared their paper for publication in PNAS, they stated that: "Because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell's Human Research Protection Program." This statement has since been confirmed by Cornell University.

Obtaining informed consent and allowing participants to opt out are best practices in most instances under the US Department of Health and Human Services Policy for the Protection of Human Research Subjects (the "Common Rule"). Adherence to the Common Rule is PNAS policy, but as a private company Facebook was under no obligation to conform to the provisions of the Common Rule when it collected the data used by the authors, and the Common Rule does not preclude their use of the data. Based on the information provided by the authors, PNAS editors deemed it appropriate to publish the paper. It is nevertheless a matter of concern that the collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out.

The Facebook Mood Experiment

Note: based on Facebook's terms of use, this was probably legal.
The question is whether it was ethical.

PNA's own editorial statement of concern summarized the issues very well

*Questions have been raised about the principles of **informed consent** and opportunity to **opt out** in connection with the research in this paper. The authors noted in their paper, "[The work] was consistent with **Facebook's Data Use Policy**, to which all users agree prior to creating an account on Facebook, **constituting informed consent for this research**." When the authors prepared their paper for publication in PNAS, they stated that: "Because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell's Human Research Protection Program." This statement has since been confirmed by Cornell University.*

*Obtaining informed consent and allowing participants to opt out are best practices in most instances under the US Department of Health and Human Services Policy for the Protection of Human Research Subjects (the "Common Rule"). Adherence to the Common Rule is PNAS policy, but as a private company Facebook was under no obligation to conform to the provisions of the Common Rule when it collected the data used by the authors, and the Common Rule does not preclude their use of the data. **Based on the information provided by the authors, PNAS editors deemed it appropriate to publish the paper. It is nevertheless a matter of concern that the collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out.***

The Facebook Mood Experiment

- What level of human experimentation is ethical?
- Is A/B testing ethical?
  - A/B testing shows different audiences different content, e.g., ads, to see which is more effective.
- What level of human manipulation is ethical?
- Is targeted advertising ethical?
- Is targeted political advertising using false information ethical?

Some Topics
From Speculative to the Here & Now

- Robot Rights
  - Humane treatment of AIs?
- Unemployment & Inequality Arising from Introduction of AI
  - How do we deal with people losing jobs
- The Ethics of Autonomous Weapons
  - *Humans in the loop* vs *Humans on the loop*
- The Ethics of Experimenting on Humans
  - Facebook's emotion experiment
- Autonomous Vehicles (self driving cars)
  - The ethics that need to be programmed into the system
- Fairness in Machine Learning
  - ML easily reinforces biases

Autonomous Vehicles (self driving cars)

- March 18, 2018, Tempe Arizona, USA
- Woman walking a Bicycle killed by UBER (HOTL) car
  - First known (?) AV fatality
  - Car had assumed that woman was another car that would move away; woman responded the way a normal bicyclist would in presence of car and didn't move.
  - To confuse issue, car handed control back to oversight driver, but very late.
- Who was responsible?
  - Tort law requires proof of intention or negligence
  - Not just ethical issue, but legal one as well
  - Uber settled with driver's family 10 days later so this case won't provide precedent

Autonomous Vehicles (self driving cars)

- 2017: Germany proposed the first (so far only) set of (20) Ethical Rules for AVs.
- Point 15:
  In the case of automated and connected driving systems, the accountability that was previously the sole preserve of the individual shifts from the motorist to the manufacturers and operators of the technological systems and to the bodies responsible for taking infrastructure, policy and legal decisions.
- *Point 7:*
  Unambiguously states that in dilemma situations, the protection of human life should enjoy top priority over the protection of other animal life.
- Point 9:
  In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited.

Autonomous Vehicles (self driving cars)

- 2017: Germany proposed the first (so far only) set of (20) Ethical Rules for AVs.

<u>Two Takeaways</u>

1. We need to encode ethical decisions into the software
   - Is this even feasible?
   - How close to perfect does the software encoding need to be before allowing it on the road?
2. German guidelines are implicitly endorsing ethical biases
   - Software not permitted to prioritize who to save
   - Parallel to this, Mercedes explicitly decided that it will prioritize passengers over pedestrians
     - "If you know you can save at least one person, at least save that one. Save the one in the car," Mercedes executive

The Trolley Problem

- Simple version (slightly different than original 50 year old one)
  see Wikipedia entry

You see a runaway trolley moving toward five tied-up people lying on the tracks.
You are standing next to a lever that controls a switch.
If you pull the lever, the trolley will be redirected onto a side track and the five people on the main track will be saved. However, there is a single person lying on the side track. You have two options:

1. Do nothing and allow the trolley to kill the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person.

Which is the more ethical option?

Problem was designed to illustrate the dilemma of whether or not
it is right to actively inhibit the utility of an individual
if doing so produces a greater utility for other individuals.

The Trolley Problem and AVs

You see a runaway trolley moving toward five tied-up people lying on the tracks.
You are standing next to a lever that controls a switch.
If you pull the lever, the trolley will be redirected onto a side track and the five people on the main track will be saved. However, there is a single person lying on the side track. You have two options:

1. Do nothing and allow the trolley to kill the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person.

Which is the more ethical option?

Trolley problem variations have become a popular way of framing ethical conflicts that can arise in AVs (which are the trolleys)

- Valuing the life of AV passenger vs. Pedestrians
  - Swerving can save the passenger but kill pedestrians
- Contrasting value of lives of different pedestrians
  - Whichever way the car is going to go, it will kill someone. Should it kill 2 old people to save one child?

The Trolley Problem and AVs

Moral Machine Project at MIT
Results reported *in Nature*, v. 563, pages 59–64 (2018)

From the web page:

We show you moral dilemmas, where a driverless car must choose between the lesser of two evils, such as killing two passengers or five pedestrians. As an outside observer, you judge which outcome you think is more acceptable.

The Trolley Problem and AVs

Moral Machine Project at MIT

Results reported *in Nature*, v. 563, pages 59–64 (2018)

From the web page:

We show you moral dilemmas, where a driverless car must choose between the lesser of two evils, such as killing two passengers or five pedestrians. As an outside observer, you judge which outcome you think is more acceptable.

The Trolley Problem and AVs

Moral Machine Project at MIT
Results reported *in Nature*, v. 563, pages 59–64 (2018)

From the web page:

We show you moral dilemmas, where a driverless car must choose between the lesser of two evils, such as killing two passengers or five pedestrians. As an outside observer, you judge which outcome you think is more acceptable.

From the Nature abstract:

….. the strongest preferences are observed for sparing humans over animals, sparing more lives, and sparing young lives.
Accordingly, these three preferences may be considered essential building blocks for machine ethics.

Do you agree with these conclusions?

The Ethics of Autonomous Vehicles

Summing Up our discussion.

- Ethical responsibility for accidents will belong to AV manufacturer
  - Q: Manufacturer of which component?
- AV software needs to encode ethical rules on how to handle accidents.
  - Q: Is this technologically possible?
- Ethical rules need to be agreed upon
  - Q: Which ethical rules?
    Might be different across cultures

The Ethics of Autonomous Vehicles

Assume that questions on previous pages have all been resolved.

A 2016 RAND institute report ran some numbers and raised an interesting question about the ethics of statistical certainty

- Assumption: before allowing fully autonomous AVs on the road, we will require clear statistical evidence that AVS are as safe as human drivers.

- Fatal crashes are uncommon.
  In 2014 according to insurance company data, there were 1.08 deaths per 100 million vehicle miles.

- “To demonstrate that fully autonomous vehicles have a fatality rate of 1.09 fatalities per 100 million miles (R=99.9999989%) with a C=95% confidence level, the vehicles would have to be driven 275 million failure-free miles.”

- “With a fleet of 100 autonomous vehicles being test-driven 24 hours a day, 365 days a year at an average speed of 25 miles per hour, this would take about 12.5 years."

The Ethics of Autonomous Vehicles

Assume that questions on previous pages have all been resolved.

A 2016 RAND institute report ran some numbers and raised an interesting question about the ethics of statistical certainty

The 12.5 years assumed an impossible rate of testing.
It would more likely require 20 or 30 years of testing to get to the level of certainty needed

What type of statistical certainty would be ethically required before releasing autonomous vehicles onto the road for general use.

Note: This also raises legal issues.
Governments would need to clarify questions of legal liability for AV manufacturers based on statistical certainty.

Some Topics
From Speculative to the Here & Now

- Robot Rights
  - Humane treatment of AIs?
- Unemployment & Inequality Arising from Introduction of AI
  - How do we deal with people losing jobs
- The Ethics of Autonomous Weapons
  - *Humans in the loop* vs *Humans on the loop*
- The Ethics of Experimenting on Humans
  - Facebook's emotion experiment
- Autonomous Vehicles (self driving cars)
  - The ethics that need to be programmed into the system
- Fairness in Machine Learning
  - ML easily reinforces biases

Fairness in Machine Learning

ML has a bias problem arising from at least two sources

1. The math has difficulty separating majority and minority populations with different characteristics
2. Bias in the training data that propagates

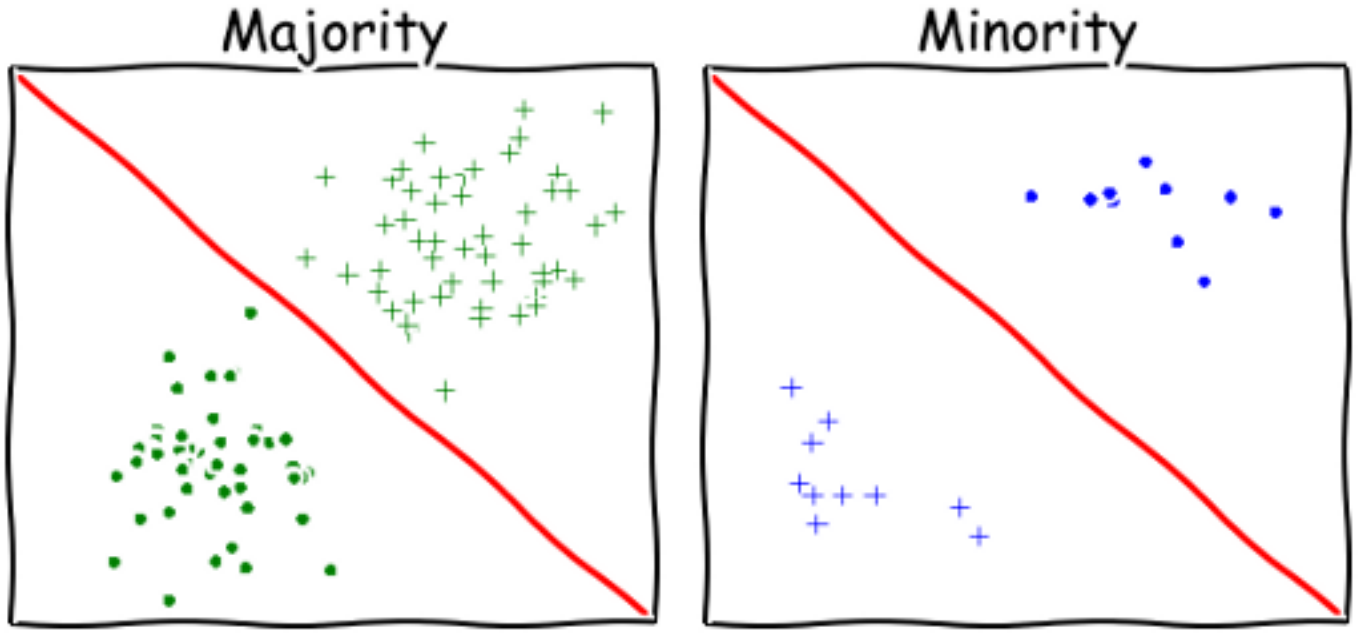
Why ML has a minority problem

- A few simple mathematical observations

- Examples and diagrams are taken from
  How big data is unfair
  Moritz Hardt
  Medium. Sept 27, 2014

- Trying to find a classifier for a property
  Population has a large majority culture with a small minority culture living alongside it.

Why ML has a minority problem (I)



Property is whether name is fake

Classifier is whether name is a long unique name

Recall: classifier is linear separator

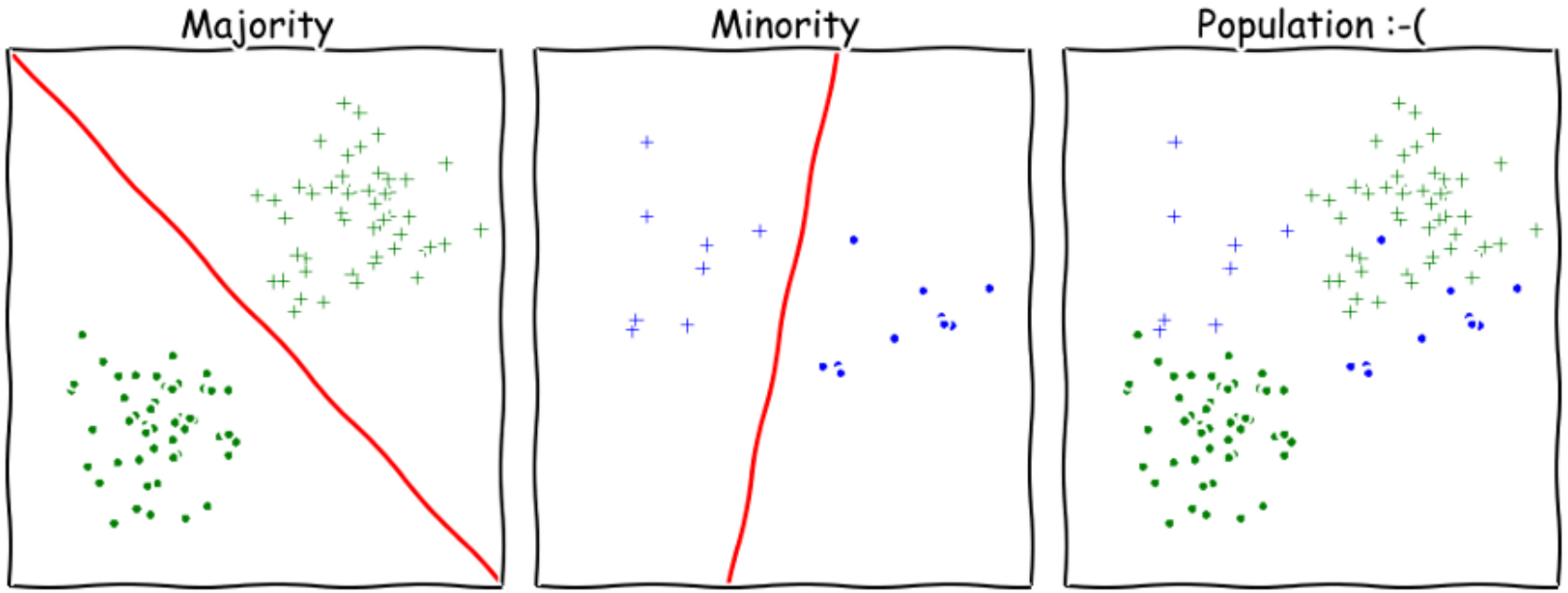
In some cultures, e.g., white American males, names are long and relatively unique

In other cultures, name can be short and repeated

Classifier that works well for majority is totally incorrect for minority!

ML wouldn't discover this because statistics show that classifier works well for most of the population

Why ML has a minority problem (II)

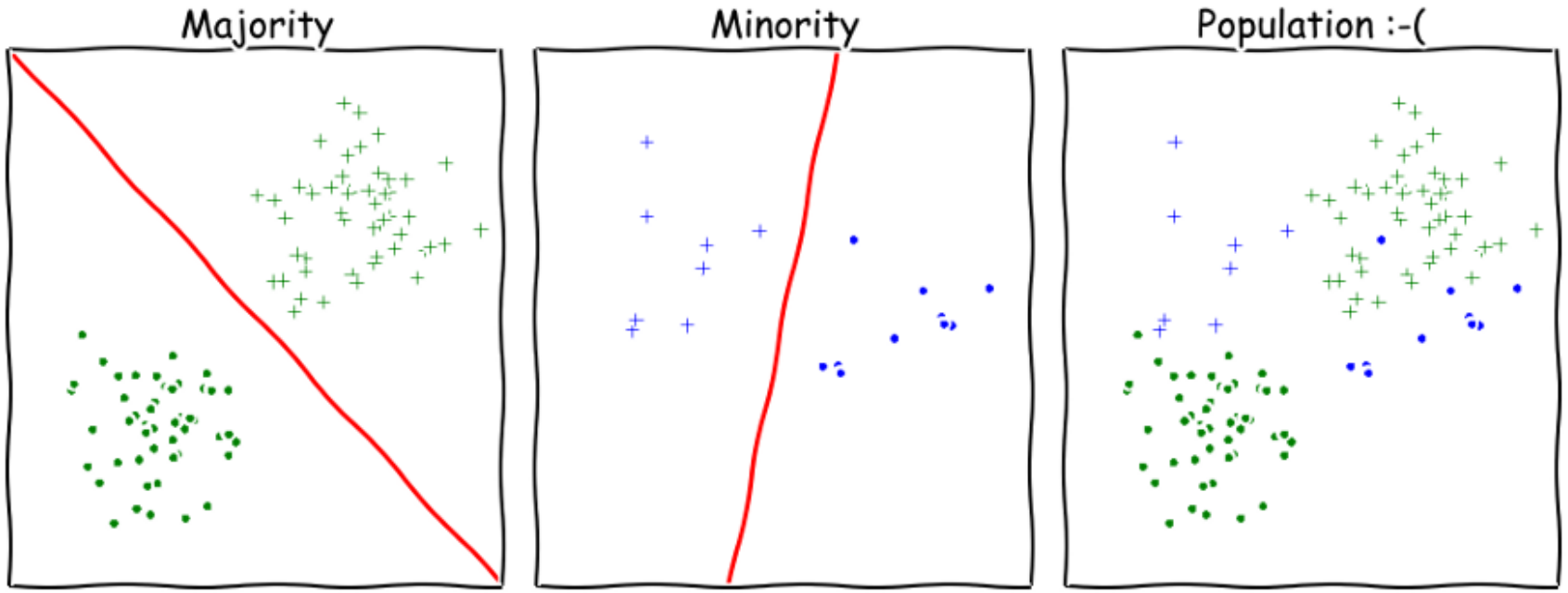


In this case both majority and minority population have good simple (linear) classifiers.

No good simple classifier exists for complete population

This embodies a deep problem at the heart of ML.
It only solves “easy” problems.

Why ML has a minority problem (II)



In this case both majority and minority population have good simple (linear) classifiers.

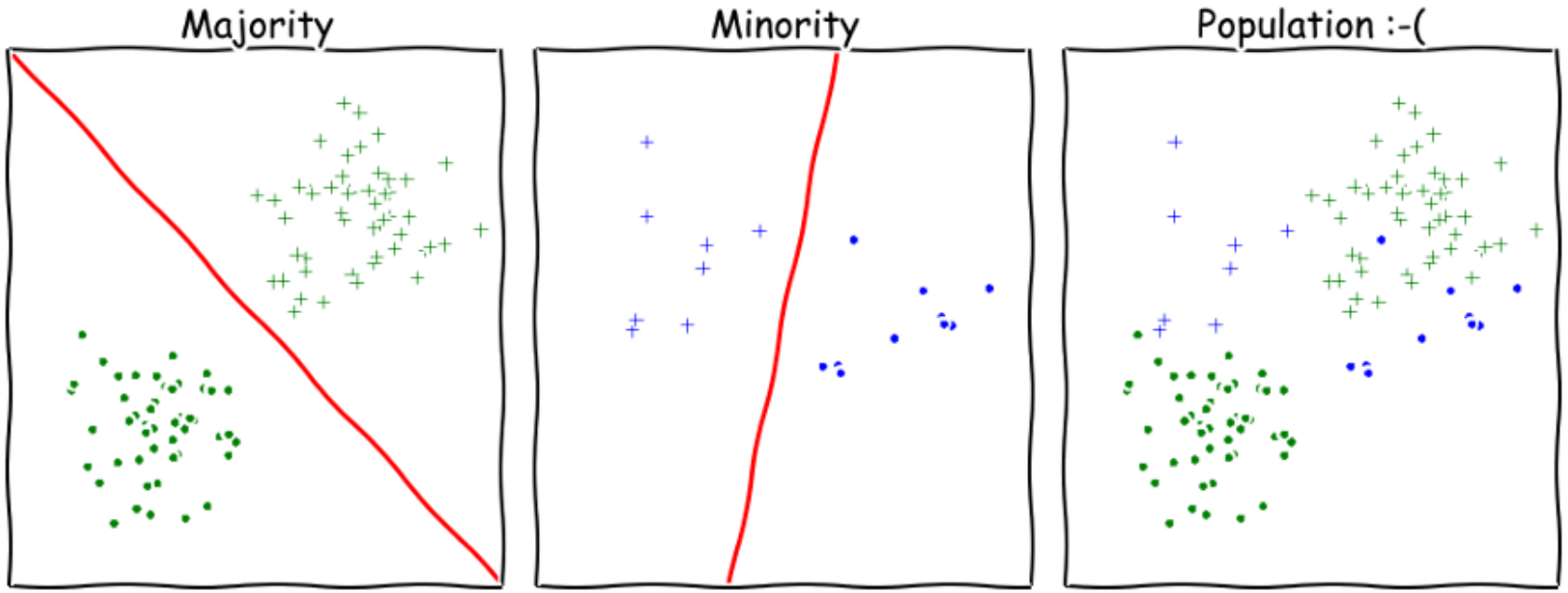
No good simple classifier exists for complete population

A classifier for the general population obviously exists.

It would first identify which population the sample belongs to and then run the appropriate population classifier on it.

This is just a combination of the two classifiers.

Why ML has a minority problem (II)



In this case both majority and minority population have good simple (linear) classifiers.

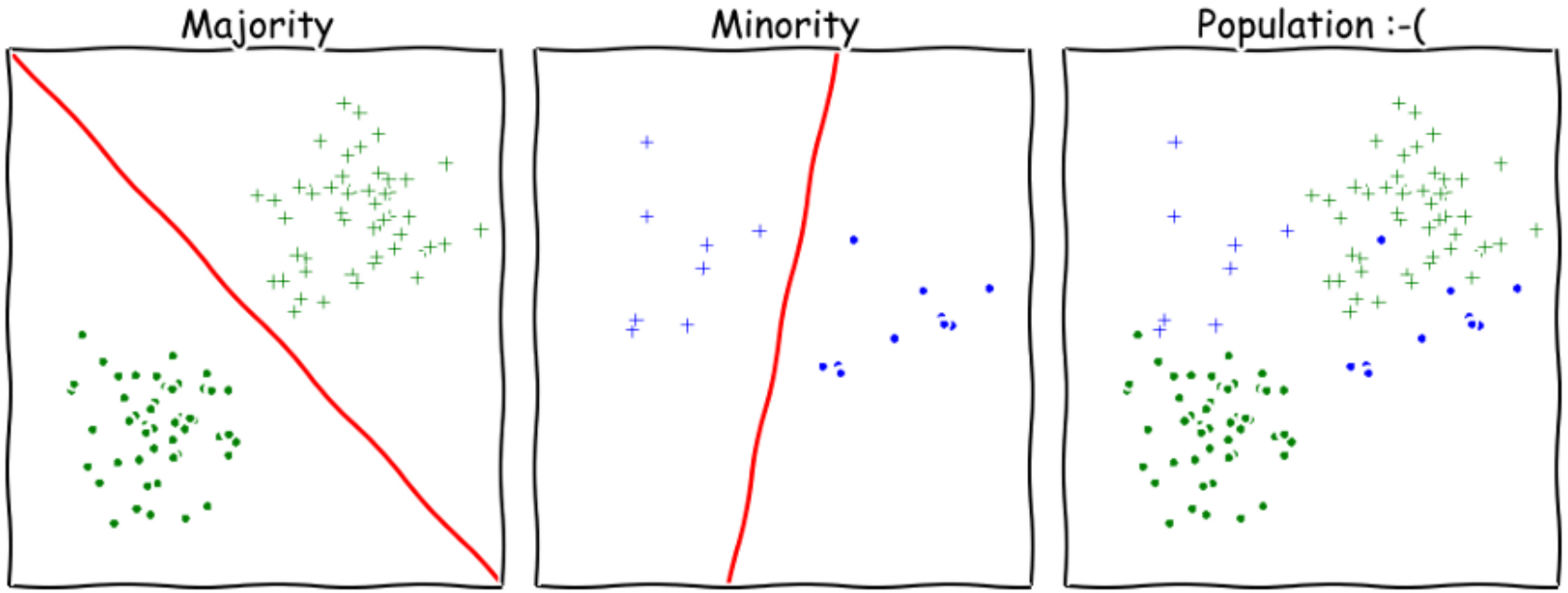
No good simple classifier exists for complete population

ML can only learn simple (linear) classifiers efficiently

We don't have good *(= computationally efficient)* techniques for learning more complicated classifiers, such as the combination one described.

So current ML technology would not be able to classify this properly

Why ML has a minority problem (II)



In this case both majority and minority population have good simple (linear) classifiers.

No good simple classifier exists for complete population

Instead, it would (probably) identify a good classifier for the majority population and leave the minority population as errors

Misclassification of minorities can have bad effects (for the minorities).

Fairness in Machine Learning

Consider the following thought experiment

- You run HR for a multi-billion dollar tech multinational employing over half a million people
- Your office is drowning in resumes
- Your tech people develop a ML based AI that, by looking at old resumes, learns the characteristics for successful resumes.
  - It will automatically throw away, say, 95% of the resumes, and leave the top 5% for you to review manually
  - After Training & Testing they report that it's fit for use

Fairness in Machine Learning

- For a while, *Resume Preprocessing Software* seems to be working well.

- This is ML using CNNs so you don't know WHAT criteria are actually being correlated for, but you don't see many false positives.

- BUT THEN!! Your office figures out that the
  - AI is ignoring relevant skills like programming ability (which is ubiquitous across resumes)
  - Main correlation is a **negative correlation against women**
    - Penalized resumes that includes "women's"
    - Downgraded graduates of all women's colleges

What do you do now!

Fairness in Machine Learning

- AI is ignoring items like programming ability (which is ubiquitous across resumes)
- Main correlation is a negative correlation against women
  - Penalized resumes that includes “women’s”
  - Downgraded graduates of all women’s colleges

- This actually happened to Amazon a couple of years ago!
  - They got ahead of the bad publicity and quickly publically announced it, killing the program

- Problem was that the training data was 10 years of Amazon resumes and the vast majority of their resumes were male. So their system picked up a high correlation between maleness and success.
  - ML learned and formalized a bias

- ML can’t tell usually you WHAT the rules it’s learned are,
  - so there was no explicit way of knowing that it had downgraded women

- Very difficult to fix it by telling it to ignore sex of candidate.
  - That feature is encoded redundantly in many correlations

Fairness in Machine Learning

- ML “learns” from training data
  - Training data is often biased
  - The classifiers and correlations it learns it can often be discriminatory, e.g., Don’t hire women. Don’t offer mortgages in minority areas.

- It can be difficult to identify this bias
  - ML is not set up to explain how its classification system works
  - This is a feature of ML not a bug
    - For classification, ML is projecting from large dimensional feature space to much lower dimensional one. No idea what the bases of these lower dimensional spaces are.
    - They can easily be encoding discrimination

- Commercial AI software often operates under Intellectual Property Rights regimens
  - so you often can’t even look under the hood to try to identify source of bias.

- Even if bias is identified, it can be almost impossible to remove
  - Because bias is correlated with many other features

Case study 2 : Machine Learning of Corpora is Biased

Caliskan A, Bryson JJ, Narayanan A.
Semantics derived automatically from language corpora contain human-like biases
Science. 2017 Apr 14;356(6334):183-6.

"**We show that standard machine learning can acquire stereotyped biases from textual data that reflect everyday human culture**."

"In another widely publicized study, Bertrand and Mullainathan (7) sent nearly 5000 identical résumés in response to 1300 job advertisements, varying only the names of the candidates. They found that European-American candidates were 50% more likely to be offered an opportunity to be interviewed. …..
We provide additional evidence for this hypothesis using word embeddings ….
We confirmed the association using two different sets of "pleasant/unpleasant" stimuli."

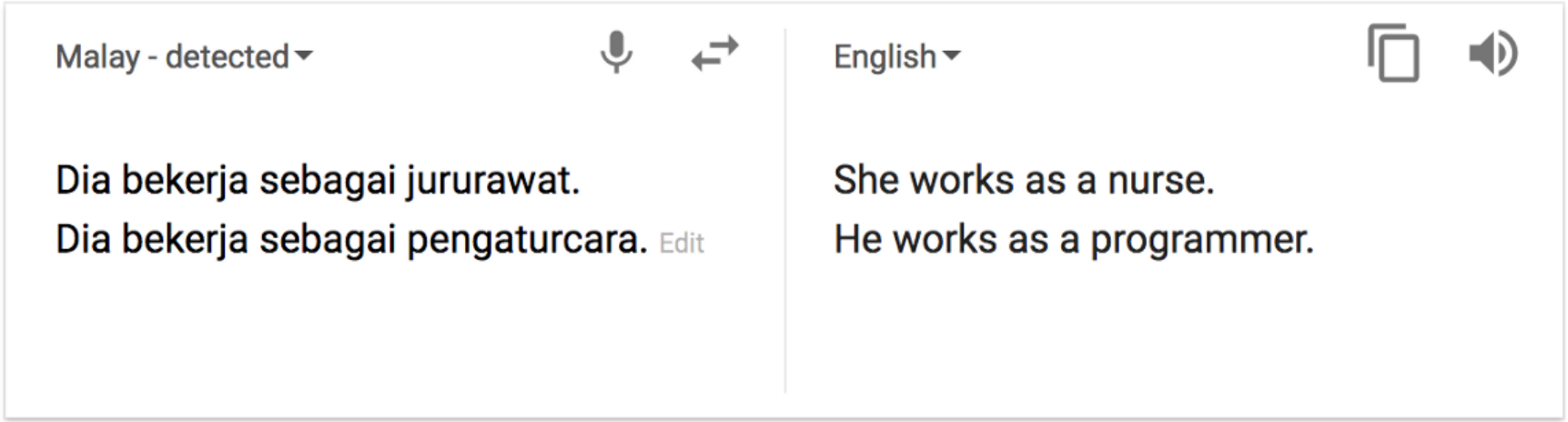
"Turning to gender biases, we replicated a finding that female names are more associated with family than career words, compared with male names"

Their research was done using off the shelf ML software, implying that this type of bias would be pervasive in applications.

Case study 3: Machine Translation is Biased

Similarly, researchers have noted noted that this bias can effect machine translation from gender neutral languages:

Example from https://hackernoon.com/bias-sexist-or-this-is-the-way-it-should-be-ce1f7c8c683c



Not unexpected when using statistical ML, but still disturbing to see so explicitly.

Case Study 4: Risk Assessment for Criminal Recidivism

- In the United States there are computer programs that do risk assessment, trying to predict whether a convicted criminal will commit a crime again
- These scores are used by judges to decide lengths of sentences to impose.

- One such program is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) sold by NORTHPOINTE

- Investigative reporting by Pro-Publica seems to show that COMPAS is race biased:
  - The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
  - White defendants were mislabeled as low risk more often than black defendants.

- COMPAS derives its score from 137 questions that are either answered by the defendants or pulled from the records. Examples are

  - *Was one of your parents ever sent to jail or prison?"*
  - *How many of your friends/acquaintances are taking drugs illegally?"*
  - *Agree or Disagree: A hungry person has a right to steal"*

  - Easy to imagine how being poor and a minority will be found by the system to correlate with high risk of recidivism even if individual would not have a high risk

  - There is no way to argue with the score
  - Even worse, scoring system is a trade secret and result can not be checked

Is This Ethical?

- COMPAS is textbook example of how NOT to design an ethical AI.
  Non-Transparent. No seeming concern for bias.
- If we went back to basics, could we design an ethical sentencing system from scratch, one that was unbiased?

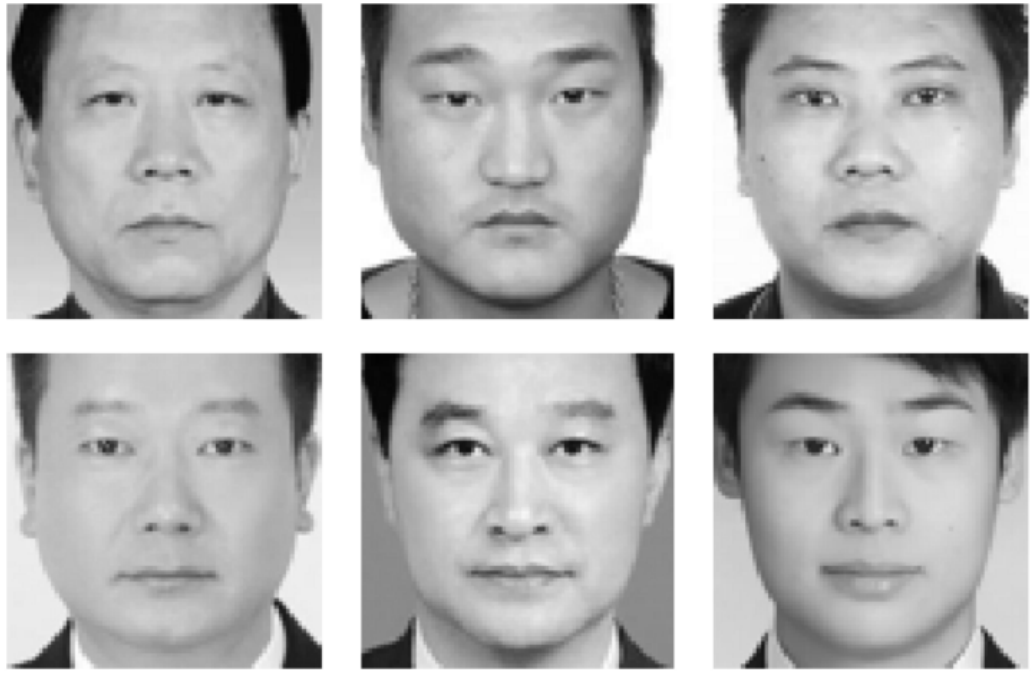
Not clear if this is possible. The US state of Pennsylvania is trying.

- Following legislation passed in 2010, Penn has been holding open transparent hearings about how the software would work and what inputs it would use.
- Even their critics agree that this is a good faith process following best practices.
- Example: Public defenders organizations argued that use of past *arrest records* would likely exacerbate racial bias (why?). Commission agreed and decided to switch to *conviction records*.
- Many people still question whether a "clean" algorithm is possible because a lot of the "objective" data is biased
  - Example: Known fact that "black people are arrested for marijuana possession at a much higher rate than white people, though they use the drug at an equal rate."
- If training data is biased then classification system based on that data will institutionalize bias and make it even harder to fight against.
- Penn has postponed vote on use of their new system to hold more hearings and evaluations ….

Perhaps no Ethical AI for this problem is possible!

Case Study 5: Modern Phrenology

Wu and Xi, Automated inference on criminality using face images, 2017



Trained on around ~1800 ID photos
730 were “criminals” (of which 330 were wanted suspects)

Authors wanted to help with predictive policing.

Claimed that their ML machine techniques can predict the likelihood that a person is a convicted criminal with nearly 90% accuracy using nothing but a driver’s license-style face photo

Method did not correct for existing biases in the justice system.
E.g., who justice system is likely to indict/convict
Also had issues with overfitting data (size of data set)

Note: paper was never published in peer-reviewed journal but did get a lot of press

Case Study 6: Predictive Policing

- Systems such as Predpol and Command Central look at current crime data and “learn” where and when crimes are more likely to occur
- Police departments then send more patrols to those locations.
- Does seem to lead to more arrests. Successful.
- Is this biased?

Algorithm is not biased but initial data likely is.

1) Poor minority areas tend to have more explicit crime as recorded by arrest records
2) => Predpol sends most of the police patrols to those neighborhoods
3) => More people arrested in those neighborhoods
4) => Back to (1), reinforcing police scrutiny of those areas.
5) Meanwhile, nonviolent crimes in more affluent neighborhoods get ignored

Case Study 7:Hate Speech Detection

- *Perspective* is a tool from Jigsaw/Alphabet that uses a convolutional neural network to detect toxic language, i.e, Hate Speech
- It's trained on crowdsourced data (tweets), which annotators were asked to label as offensive or not (toxicity).
- Public Service to help automate hate speech detection

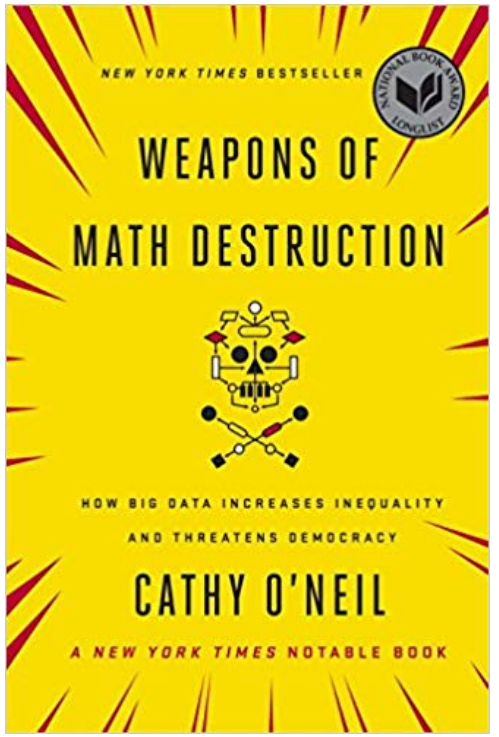
What could possibly go wrong?

Case Study 7:Hate Speech Detection

- Uncovered correlations between use of African American English (AAE) dialect and toxicity rating.
  - This means that same idea expressed in AAE was twice as likely to be labelled as offensive as if not in AAE.
- Remember that the labelling of the training data was crowdsourced.
  - Unsurprisingly, the crowdsourced labelling reflected a social bias against AAE
  - This bias was learned by the CNN
- Good news is that when labellers were made explicitly aware of tweeter's dialect bias was significantly reduced (tweets were not labelled offensive).

The Risk of Racial Bias in Hate Speech Detection
M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith
Proc. of the 57th Annual Meeting of the Association for Computational Linguistics
July 28-Aug 2, 2019

Data Scientist Dr. Cathy O'Neil calls some of the algorithms we have seen:
Weapons of Math Destruction



Book lists many other examples

Three Key Features:

1. Algorithm is opaque
   - No idea WHY it makes the decisions it does
     Standard for ML algorithms
   - Opaque and invisible models are the rule and clear ones very much the exception
   - Often provided by companies and not by government
     Algorithm internals are an IP secret

2. Has negative impact on individual's life
   - Denying credit based on race
   - Resumes automatically ignored
   - Longer jail sentences based on correlations that would be illegal if they were explicit

3. (Massively) Scalable
   - Can easily grow exponentially. Becomes more powerful as it scales
   - If a bank's credit model scales to the country, it affects your whole life.
   - If Recidivism score is used by more counties, it can destroy more lives

What can be done to alleviate bias?

- Research Side
  - New emerging field of machine learning fairness
    - Much harder and techniques not nearly as efficient as standard ML
  - New research into ML algorithms that explain HOW their classifiers work
    - Same issues
      Much harder and techniques not nearly as efficient as standard ML
  - Will take time and unclear how successful they will be

What can be done to alleviate bias?

- User Side: Industry and Government
  - Governments might need to institute anti-discriminatory regulation
    - Already happening.
      New York City passed an AI Transparency Bill in January 2018 and set up an Automated Decision Systems Task Force to review algorithms used by the city government.
  - Industry
    - New Partnership on AI
      Many major corporations: Facebook, Google, Baidu, Amazon, Accenture, McKinsey
      Many NGOS: ACLU, Human Rights Watch, Berkeley Center for Law
    - Working for fair accountable AI

Will they succeed?

Unknown. But at least the issue has been recognized pretty early on.

Principles for Accountable Algorithms *(from FAT-ML)*

- **Responsibility**
  - Make available externally visible avenues of redress for adverse individual or societal effects of an algorithmic decision system, and designate an internal role for the person who is responsible for the timely remedy of such issues.
- **Explainability**
  - Ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms.
- **Accuracy**
  - Identify, log, and articulate sources of error and uncertainty throughout the algorithm and its data sources so that expected and worst case implications can be understood and inform mitigation procedures.
- **Auditability**
  - Enable interested third parties to probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of detailed documentation, technically suitable APIs, and permissive terms of use.
- **Fairness**
  - Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics (e.g. race, sex, etc).

Some Topics
From Speculative to the Here & Now

- Robot Rights
  - Humane treatment of AIs?
- Unemployment & Inequality Arising from Introduction of AI
  - How do we deal with people losing jobs
- The Ethics of Autonomous Weapons
  - *Humans in the loop* vs *Humans on the loop*
- The Ethics of Experimenting on Humans
  - Facebook's emotion experiment
- Autonomous Vehicles (self driving cars)
  - The ethics that need to be programmed into the system
- Fairness in Machine Learning
  - ML easily reinforces biases

The Ethics of Privacy

This is a huge topic. Its own lecture, if not its own course. A few starting points

- **Privacy as a Human Right**
  - 1948: Universal Declaration of Human Rights
- **What is the responsibility of AI practitioners to maintain those rights?**
  - Guide your data well
  - Be aware that it is extremely hard to anonymize data, especially with powerful social network analysis techniques. See Netflix Challenge.
- **Best practices**
  - Only collect as much data is necessary
  - Only release as much data as necessary
    - Differential privacy
      mathematical definition of how much privacy loss comes from aggregate release of data
    - Add noise to data before release

Very Hard to Keep Data Private/Anonymous

- 2006 Netflix Prize Case Study
  - Netflix released a anonymized dataset of more than 450,000 movie recommendations made by their customers
  - They were offering a 1 million dollar prize to best recommendation algorithm (that could be trained on this data)
  - Netflix claimed that “all customer identifying information has been removed”
  - Researchers were able to identify many of the customers by correlating with IMDB public ratings!

Very Hard to Keep Data Private/Anonymous

- 2006 Netflix Prize Case Study
  - Netflix released a anonymized dataset of more than 450,000 movie recommendations made by their customers
- Much work since then
  - All showing that it is almost impossible to anonymize data sets

- 2019 Luc Rocher, Julien M. Hendrick and Yves-Alexandre de Montjoye
  Estimating the success of re-identifications in incomplete datasets using generative models
  Nature Communications volume 10, (2019)
  - *"Using our model, we find that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes."*
  - *"Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model."*

Some last thoughts

- This talk has just scratched the surface
- I hope that it at least got you thinking
- AI and especially the combination of ML with Big Data is an extraordinarily powerful tool
  - Like most powerful tools it can be used for both good and bad purposes
  - It does raise more ethical quandaries than most
  - One reason for this is that it has become so powerful so fast (confluence of multiple factors)
  - Haven't had time to consider the consequences before employing
  - Once one actor employs, others feel the necessity to employ as well
- Many other issues that we didn't address
  - One big one is the loss of privacy as a human right
- For a more detailed look at this topic see the extra reading list
- Pointers to more than 100 syllabi on ethics and big data