Societal consequences
A few vignettes
A few vignettes

Machine learning is entering human-facing technologies more and more.

- “Big data” and large scale machine learning allow for mass surveillance (and mass discrimination).
- The choices suggested by machine learning systems are often inscrutable and unreliable; the systems are hard to debug.
- Machine learning systems have obscure failure modes (adversarial examples!) which well-informed attackers can trigger.
A few vignettes

Machine learning is entering human-facing technologies more and more.

▶ “Big data” and large scale machine learning allow for mass surveillance (and mass discrimination).

▶ The choices suggested by machine learning systems are often inscrutable and unreliable; the systems are hard to debug.

▶ Machine learning systems have obscure failure modes (adversarial examples!) which well-informed attackers can trigger.

How can anyone argue a machine learning system actually implements some policy? How can one argue it doesn’t covertly implement some other policy?
The music industry has criticized Content ID as inefficient, with Universal Music Publishing Group (UMPG) estimating in a 2015 filing to the US Copyright Office "that Content ID fails to identify upwards of 40 percent of the use of UMPG’s compositions on YouTube". [1][27] Google has countered these assertions by stating that (as of 2016) Content ID detected over 98% of known copyright infringement on YouTube and humans filing removal notices only 2%. [1]

In January 2018, a YouTube uploader who created a white noise generator received copyright notices about a video he uploaded which was created using this tool, and therefore containing only white noise. [28]

In September 2018, a German university professor uploaded videos with several classical music performances for which their copyright had expired, because both the composers were dead long ago, and the performances were not covered anymore by copyright. After he received several copyright violations by YouTube, he could lift the majority of them, but Deutsche Grammophon refused to lift two of them even if their copyright had expired. [29][30][31] In other cases, copyright violations notices were even sent to uploaders who recorded themselves playing public domain classical music, with Sony Music asserting copyright over more than 1,100 compositions by Johann Sebastian Bach via Content ID. [32] Commentators noted that this was also the case on other platforms such as Facebook. [33]

In December 2018 TheFatRat complained that Content ID gave preference to an obvious scammer who used the automated system to claim ownership of his content and thereby steal his revenue. [34]

(From Wikipedia.)

YouTube can label videos as violating copyright and take them down. This can remove livelihood (demonetization).
China’s “social credit system”

Social Credit System

From Wikipedia, the free encyclopedia

For the economic and political philosophy founded by British engineer C. H. Douglas (1879–1952), see Social credit.

The Social Credit System (Chinese: 社会信用体系; pinyin: shèhuì xìnyòng tǐxì) is a national reputation system being developed by the Chinese government.[1][2][3] By 2020, it is intended to standardise the assessment of citizens' and businesses' economic and social reputation, or 'Social Credit'.[4][5][6][7][8]

The system will be one unified system and there will be a single system-wide social credit score for each citizen and business.[9] By 2018, some restrictions had been placed on citizens, which state-owned media described as the first step toward creating a national social credit system.[10][11][12][7][13][14]

The system is considered a form of mass surveillance which uses big data analysis technology.[15] The government of modern China has also maintained systems of paper records on individuals and households such as the dàngàn (档案) and hùkǒu (户口) systems which officials might refer to, but did not provide the same degree and rapidity of feedback and consequences for Chinese citizens as the integrated electronic system because of the much greater difficulty of aggregating paper records for rapid, robust analysis.

(From Wikipedia.)

Who oversees this system?
Which political agendas does it actually enforce?
Other examples

- What if everyone starts using deep-learning based websites to find dates?
  - What if someone tampers with the algorithms?
- What if all cars are self-driving, with deep-learning based vision systems?
  - What if someone tampers with the weight vectors so that, in rare situations, the car drives itself into oncoming traffic? Given the existence of adversarial examples, are such modifications detectable?
An analysis of Juniper's ScreenOS firmware code in December 2015 discovered a backdoor key using Dual_EC_DRBG allowing to passively decrypt the traffic encrypted by ScreenOS. This backdoor was inserted in the year 2008 into the versions of ScreenOS from 6.2.0r15 to 6.2.0r18 and from 6.3.0r12 to 6.3.0r20[145] and gives any user administrative access when using a special master password.[146] Some analysts claim that this backdoor still exists in ScreenOS.[147] Stephen Checkoway was quoted in Wired that "If this backdoor was not intentional, then, in my opinion, it’s an amazing coincidence."[148]

In December 2015, Juniper Systems announced that they had discovered "unauthorized code" in the ScreenOS software that underlies their NetScreen devices, present from 2012 onwards. There were two vulnerabilities: One was a simple root password backdoor, and the other one was changing a point in Dual_EC_DRBG so that the attackers presumably had the key to use the preexisting (intentional or unintentional) kleptographic backdoor in ScreenOS to passively decrypt traffic.[149]

(From Wikipedia.)
Once nation-states start tampering with machine learning systems, it’ll get even scarier.
Facial recognition and other biometrics?

Facial recognition creeps up on a JetBlue passenger and she hates it

Facial recognition systems don’t want you to stop and think about them. This is what happens when someone does.

By Chris Miyazaki for Technically Incorrect | April 22, 2009 -- 16:42 CMT (06:42 PDT) Topic: Security

Naturally, she took to Twitter to register her troubles.

She began: "I just boarded an international @JetBlue flight. Instead of scanning my boarding pass or handing over my passport, I looked into a camera before being allowed down the jet bridge. Did facial recognition replace boarding passes, unbeknownst to me? Did I consent to this?"

A funny thing, consent. Sometimes, you have no idea you’ve already given it. Sometimes, you haven’t given it at all.

JetBlue, as all good airlines do, was ready to offer Twitterized sympathy: "You’re able to opt out of this procedure. Mackenzie. Sorry if this made you feel uncomfortable.”

But once you start thinking about these things, your thoughts become darker. Fegan wanted to know how JetBlue knew what she looked like.

JetBlue explained: "The information is provided by the United States Department of Homeland Security from existing holdings.'

Fegan wondered by what right a private company suddenly had her biometric data.

JetBlue insisted it doesn’t have access to the data. It’s "securely transmitted to the Customs and Border Protection database.”

(From ZDNet.)
Facial recognition

Color Matters in Computer Vision
Facial recognition algorithms made by Microsoft, IBM and Face++ were more likely to misidentify the gender of black women than white men.

Gender was misidentified in up to 1 percent of lighter-skinned males in a set of 318 photos.

Gender was misidentified in up to 7 percent of lighter-skinned females in a set of 296 photos.

Gender was misidentified in up to 12 percent of darker-skinned males in a set of 318 photos.

Gender was misidentified in 35 percent of darker-skinned females in a set of 271 photos.

(From NYTimes.)
How to hide from the AI surveillance state with a color printout

AI-powered video technology is becoming ubiquitous, tracking our faces and bodies through stores, offices, and public spaces. In some countries the technology constitutes a powerful new layer of policing and government surveillance.

Fortunately, as some researchers from the Belgian university KU Leuven have just shown, you can often hide from an AI video system with the aid of a simple color printout.

**Who said that?** The researchers showed that the image they designed can hide a whole person from an AI-powered computer-vision system. They demonstrated it on a popular open-source object recognition system called YoLo(v2).

**Hide and seek:** The trick could conceivably let crooks hide from security cameras, or offer dissidents a way to dodge government scrutiny. "What our work proves is that it is possible to bypass camera surveillance systems using adversarial patches," says
Today we’ll discuss two fields of study related to societal consequences: fairness and privacy. Here are some references:


These fields are still young, and most deployed ML is not adjusted to take them into account.
Privacy
Suppose each example in a data set $D = (z_i)_{i=1}^n$ corresponds to an individual from the population.

▶ Can the output of a learning algorithm, run on data set $D$, reveal information about an individual?

▶ **Example**: average salary computation.
   (Recall, average is minimizer of squared loss risk . . .)

Suppose attacker knows salary of all but one person in a company.

**Claim**: By learning the average salary, attacker knows everyone’s salary.

$$\text{average salary} = \frac{1}{n} \left( \text{your salary} + \text{sum of everyone else’s salary} \right).$$

▶ Privacy was compromised by release of average salary:
   ▶ Before release of average salary, attacker did not know your salary.
   ▶ After release of average salary, attacker knows your salary.
Privacy compromised by release of genome-wide association study (GWAS) results (Wang, Li, Wang, Tang, & Zhou, 2009).

“individuals can be identified from even a relatively small set of statistics, as those routinely published in GWAS papers”

- Based on your genes, can learn if you participated in the study or not.
- Can also learn if you were in the “disease” group or “healthy” group.
Scenario: An attacker knows your genetic information, but does not know if you have certain disease.

1. Suppose GWAS results allow attacker to determine that you were in the “disease” group in the study.
   Is privacy compromised by release of GWAS results?

2. Suppose GWAS results suggest people (like you) with a certain genetic marker are likely to have a disease.
   Is privacy compromised by release of GWAS results?
What we would like:
Given output of learning algorithm, attacker should not be able to distinguish between:

- **Case 1**: Training data contains your data \((x_i, y_i)\).
- **Case 2**: Training data does not contain your data \((x_i, y_i)\).

- Whether or not you are included in training data does not change what attacker can learn about you.
- Any secrets you have (that couldn’t have been learned otherwise) are not compromised by the learning algorithm.

Learning algorithms with this property are said to guarantee **differential privacy** (Dwork, McSherry, Nissim, & Smith, 2006).
**Formal definition:** A randomized algorithm $\mathcal{A}$ guarantees $\epsilon$-differential privacy if, for all data sets $D$ and $D'$ that differ in just one individual's data point,

$$\mathbb{P}(\mathcal{A}(D) = t) \in (1 \pm \epsilon) \cdot \mathbb{P}(\mathcal{A}(D') = t), \quad \forall \text{ possible outputs } t.$$
Example: differentially private average of salaries

**Average salary**

- **Goal**: approximately compute average salary.
Example: differentially private average of salaries

**Average salary**

- **Goal**: approximately compute average salary.
- Sample average is not differentially private as before: can use it to recover individual salaries.

\[ A(D) = \frac{1}{n} \sum_{i=1}^{n} s_i + Z, \]

where \( Z \sim p_Z(z) = \frac{1}{2\sigma} \exp(-|z|\sigma) \), \( \sigma \approx \text{max salary} / \epsilon n \).

("max salary" is assumed to be publicly known.)
Example: differentially private average of salaries

Average salary

- **Goal**: approximately compute average salary.
- Sample average is not differentially private as before: can use it to recover individual salaries.

- **Our algorithm**: Given data set $D = (s_i)_{i=1}^n$ of salaries, output

$$A(D) = \frac{1}{n} \sum_{i=1}^n s_i + Z,$$

where

$$Z \sim p_Z(z) = \frac{1}{2\sigma} \exp \left( -\frac{|z|}{\sigma} \right),$$

$$\sigma \approx \frac{\text{max salary}}{\epsilon n}.$$

("max salary" is assumed to be publicly known.)
Why this guarantees differential privacy

Consider data sets of salaries $D$ and $D'$ that differ only in one entry.

- **Our algorithm:** $A(D) = \text{avg}(D) + Z$, where
  
  $$Z \sim p_Z(z) = \frac{1}{2\sigma} \exp \left( -\frac{|z|}{\sigma} \right), \quad \sigma \approx \frac{\text{max salary}}{\epsilon n}. $$

- Need to show that output densities of $A$ on input $D$ and on input $D'$ are within $1 \pm \epsilon$ of each other.

- Since $D$ and $D'$ differ in one entry, say, $s_i$ and $s'_i$:
  
  $$A(D') = \text{avg}(D) + \frac{s'_i - s_i}{n} + Z.$$

- Ratio of probability densities:
  
  $$\frac{p_{A(D)}(t)}{p_{A(D')}(t)} = \frac{p_Z(t - \text{avg}(D))}{p_Z(t - \text{avg}(D) - \frac{s'_i - s_i}{n})}$$

  $$= \exp \left( -\frac{1}{\sigma} \left( |t - \text{avg}(D)| - |t - \text{avg}(D) - \frac{s'_i - s_i}{n}| \right) \right)$$

  $$\in \exp \left( \pm \frac{1}{\sigma} \cdot \frac{\text{max salary}}{n} \right) \approx 1 \pm \epsilon.$$
For statistical analysis: Often data set itself is just a random sample, and we care more about broader population.

- Average on sample differs from population mean by $\Theta(n^{-1/2})$ anyway!

- Expected magnitude of noise $Z$: $\sigma \approx \frac{\text{max salary}}{\epsilon n}$.

Overall:

$$|\mathcal{A}(D) - \text{population mean}| \leq O \left( \frac{\text{salary stddev}}{\sqrt{n}} + \frac{\text{max salary}}{\epsilon n} \right).$$

Accuracy improves with $n$, even under constraint of $\epsilon$-differential privacy.

- In particular, when max salary is comparable to salary stddev, $\epsilon \approx 1/\sqrt{n}$ is almost “free”.
Recap: privacy

- Privacy can be compromised when sensitive data is used in data analysis / machine learning.
- Is one formal definition of privacy that can be rigorously guaranteed and reasoned about.
- Many techniques for creating “differentially private” versions of statistical methods (including machine learning algorithms).
Fairness

What is fairness?

▶ Equal treatment of people.
▶ Equal treatment of groups of people.
▶ . . .

What is unfair?

▶ Unequal treatment (of people or groups of people) based on attributes that society deems inappropriate (e.g., race, sex, age, religion).
▶ . . .

**Highly domain- and application-specific.**

Why is this of concern with machine learning?

▶ Systems built using machine learning are increasingly used to make decisions that affect people’s lives.
▶ More opportunity to misuse and confuse.
Example: predictive policing

*Predictive policing*: data-driven systems for allocating police resources based on predictions of where it is needed.

Financial Times article from August 22, 2014 by Gillian Tett:

*After all, as the former CPD computer experts point out, the algorithms in themselves are neutral. “This program had absolutely nothing to do with race ... but multi-variable equations,” argues Goldstein. Meanwhile, the potential benefits of predictive policing are profound.*
Unfairness with machine learning

Many reasons systems using machine learning can be unfair:

▶ People deliberately seeking to be unfair.
▶ Disparity in amounts of available training data for different individuals/groups.
▶ Differences in usefulness of available features for different individuals/groups.
▶ Differences in relevance of prediction problem for different individuals/groups.
▶ . . .

How we might address these problems:

1. Rigorously define and quantify (un)fairness.
2. Identify sources of confusion regarding fairness.
3. . . .
Quantifying (un)fairness: statistical parity

Setup for classification problem:

- **X**: features of an individual.
- **A**: protected attribute (e.g., race, sex, age, religion).
  - For simplicity, assume A takes on only two values, 0 and 1.
- **Y**: output variable to predict (e.g., will repay loan).
  - For simplicity, assume Y takes on only two values, 0 and 1.
- **Ŷ**: prediction of output variable (as function of X and A).

*Statistical parity* is satisfied if

\[
P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)
\]

(i.e., \(\hat{Y}\) is independent of A).

A particular relaxation is the **4/5th rule**:

\[
P(\hat{Y} = 1 \mid A = 0) \geq \frac{4}{5} \cdot P(\hat{Y} = 1 \mid A = 1).
\]
Proxies for protected attribute

Cannot simply avoid using $A$ explicitly, since because other features could be correlated with it.

Example: WSJ observed that online price for an item depends on how far you live from a brick-and-mortar Staples store.

- Pricing system doesn’t explicitly look at your income.
- But where you live is probably correlated with your income.
- Result: charging higher prices to lower-income people.
Problem with statistical parity

The following example satisfies statistical parity:

<table>
<thead>
<tr>
<th></th>
<th>$A = 0$</th>
<th></th>
<th>$A = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{Y} = 0$</td>
<td>1/2</td>
<td>$\hat{Y} = 1$</td>
<td>0</td>
</tr>
<tr>
<td>$Y = 0$</td>
<td>1/4</td>
<td>$\hat{Y} = 0$</td>
<td>1/4</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>0</td>
<td>$\hat{Y} = 1$</td>
<td>1/4</td>
</tr>
</tbody>
</table>

Suppose $Y = 1\{\text{will repay loan if granted one}\}$.

- **For group** $A = 0$: correctly give loans to people who will repay them.
- **For group** $A = 1$: give loans out at random. (Sets up people for failure!)
Equalized odds is satisfied if
\[
P(\hat{Y} = 1 \mid Y = y, A = 0) = P(\hat{Y} = 1 \mid Y = y, A = 1), \quad y \in \{0, 1\}
\]
(i.e., \(\hat{Y}\) is conditionally independent of \(A\) given \(Y\)).

- The previous example fails to satisfy equalized odds:

<table>
<thead>
<tr>
<th></th>
<th>(A = 0)</th>
<th>(A = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{Y} = 0)</td>
<td>(\frac{1}{2})</td>
<td>(1/4)</td>
</tr>
<tr>
<td>(\hat{Y} = 1)</td>
<td>0</td>
<td>(1/4)</td>
</tr>
<tr>
<td>(Y = 0)</td>
<td>(1/2)</td>
<td>0</td>
</tr>
<tr>
<td>(Y = 1)</td>
<td>0</td>
<td>(1/4)</td>
</tr>
</tbody>
</table>

\[
\frac{0}{0+1/2} \neq \frac{1/4}{1/4+1/4}
\]

- Possible to post-process a predictor to make it satisfy equalized odds (w.r.t. empirical distribution), provided that predictor can look at \(A\). (Hardt, Price, & Srebro, 2016)

- **Problem:** may be illegal to look at \(A\), even if purpose is to ensure fairness property.
Within-group (weak) calibration is satisfied if

\[ \mathbb{P}(\hat{Y} = 1 \mid A = a) = \mathbb{P}(Y = 1 \mid A = a), \quad a \in \{0, 1\}. \]

**Note:** Suppose \( \hat{Y} \) is randomized classifier derived from conditional probability estimator \( \hat{p} \), i.e.,

\[ \hat{Y} = \begin{cases} 
1 & \text{with probability } \hat{p}, \\
0 & \text{with probability } 1 - \hat{p}.
\end{cases} \]

And further, suppose \( \hat{p} \) satisfies

\[ \mathbb{P}(Y = 1 \mid \hat{p} = p, A = 0) = \mathbb{P}(Y = 1 \mid \hat{p} = p, A = 1) = p, \quad p \in [0, 1], \]

(i.e., \( \hat{p} \) is calibrated conditional probability estimator). Then \( \hat{Y} \) satisfies within-group (weak) calibration.
Theorem (Kleinberg, Mullainathan, & Raghavan, 2016): Impossible to satisfy both within-group (weak) calibration and equalized odds, unless

1. $P(Y = 1 \mid A = 0) = P(Y = 1 \mid A = 1)$, or
2. $\hat{Y}$ is perfect predictor of $Y$.

Note: Similar dichotomy holds for scoring functions / conditional probability estimators, with slightly different definitions of within-group calibration and equalized odds.
COMPAS “risk assessment”: predict whether criminal defendants will commit (another) crime.

- ProPublica study argues COMPAS is unfair:

  ![Table]

<table>
<thead>
<tr>
<th></th>
<th>$A = 0$</th>
<th>$A = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{Y} = 0$</td>
<td>0.27</td>
<td>0.46</td>
</tr>
<tr>
<td>$\hat{Y} = 1$</td>
<td>0.22</td>
<td>0.14</td>
</tr>
<tr>
<td>$Y = 0$</td>
<td>0.14</td>
<td>0.19</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>0.37</td>
<td>0.21</td>
</tr>
</tbody>
</table>

  E.g., FPR for $A = 0$ is 0.45, while FPR for $A = 1$ is 0.23.

- Study also shows COMPAS satisfies approximate within-group (weak) calibration.

  \[
  \mathbb{P}(Y = 1 \mid \hat{Y} = 1, A = 0) = 0.64, \quad \mathbb{P}(Y = 1 \mid \hat{Y} = 1, A = 1) = 0.59.
  \]

  (This was part of a different argument that COMPAS is fair.)

- Inherent trade-off theorem explains why different definitions of fairness may necessarily lead to different conclusions.
Sometimes the available training data is not suitable for task. E.g., system to decide loan applications. Goal is to predict whether an applicant will repay a loan if granted one.

Is the following training data appropriate?

1. $x =$ features of past loan applicant, 
   $y = 1\{\text{loan application was approved}\}$.
2. $x =$ features of past loan applicant who received loan, 
   $y = 1\{\text{loan was repaid}\}$.
3. $x =$ features of past loan applicant, 
   $y = 1\{\text{loan was repaid}\}$. 
Recap: fairness

- Automated decision-making based on machine learning should be scrutinized for appropriate fairness considerations (like any other decision-making system).
- Must think hard about appropriateness of data for a given application.