CS 4803/7643 W15: Fairness, Accountability, and Transparency

Toby Farmer, Adam Obeng
Expectations

- Motivation for why these issues come up and matter
- A couple of specific examples
- Not an exhaustive listing of all the FAT* problems which have come up
- Not a definitive solution to any of them
- Guidelines of how to identify and address this type of problem
Overview

- What are we even talking about?
- Why should we care?
- What are the problems?
- What should we do about them?
What are we even talking about?
What are we even talking about?

FAT, FAT*, FATE, FATES, etc.

- Fairness
- Accountability
- Transparency
- Ethics
- Safety/Security
Why should we care?
Why care about FAT*?

View 0: We shouldn’t.
Why care about FAT*?

View 0: We shouldn’t.

a) OK, but other people care

The European Commission considers new regulations and enforcement for "high-risk" AI

Brookings Institution | 6 hours ago

The AAAI Code of Professional Ethics and Conduct

An AI professional should...

1.1 Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.
Why care about FAT*?

View 0: We shouldn’t.

a) OK, but other people care
b) Even if you’re going to be a Rick, these considerations matter to “pure science”
Why care about FAT*?

View 1: We need to do FAT after we do science

i.e. ethics as an organic banana sticker
Why care about FAT*?

View 2: FAT* concerns are *inextricable* from ML

- Technology affords and constrains
  - Technology is political
- Science and engineering construct abstractions
  - Knowledge and techne are social facts
What are the problems?
An Unconscionably Brief Overview of FAT* Problems

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

Facebook Manipulated 689,003 Users' Emotions For Science

Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use

NTSB Releases Report On 2018 Fatal Silicon Valley Tesla Autopilot Crash

Building inclusive AI at Facebook

https://tech.fb.com/building-inclusive-ai-at-facebook/
An Unconscionably Brief Overview of FAT* Problems

Just two in-depth examples:

- The Fairness Impossibility Theorems
- Gender and Word Embeddings
Example 1: The Fairness Impossibility Theorems

It is impossible for a classifier to achieve parity between groups, (if there is a difference in prevalence between the groups and the classifier is not perfect)


Example 1: The Fairness Impossibility Theorems

Classifier confusion matrix

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Prediction</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>FN</td>
<td>TN</td>
<td></td>
</tr>
</tbody>
</table>

Derived quantities:

- False Positive Rate (FPR): $\frac{FP}{FP+TN}$
- False Negative Rate (FNR): $\frac{FN}{FN+TP}$
- Positive Predictive Value (PPV): $\frac{TP}{TP+FP}$
  - Measures “test fairness” for a binary classifier:
Example 1: The Fairness Impossibility Theorems

Result:

\[ FPR = \frac{p}{1 - p} \frac{1 - PPV}{PPV} (1 - FNR) \]

(where \( p \) is the prevalence of the label in a given group)
Example 1: The Fairness Impossibility Theorems

More generally, we can state many fairness theorems based on any three quantities derived from the confusion matrix.

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>
Example 1: The Fairness Impossibility Theorems

An impossibility theorem obtains for any three measures of model performance derived (non degenerately) from the confusion matrix.

- In all cases, \( p = \frac{TP + FN}{TP + FP + TN + FN} \)

- In a system of equations with three more equations, p is determined uniquely: if groups have different prevalences, these quantities cannot be equal
Aside: The Tetrachoric Correlation Coefficient

- This problem is not unique to ML
- Knowledge covers its tracks
Aside: The Tetrachoric Correlation Coefficient

- Correlation for continuous variables was well defined
- How to define correlation for discrete variables?

Yule’s Q: \[ Q = \frac{ad - bc}{ad + bc} \]

Pearson’s Tetrachoric Coefficient of correlation:

- assume underlying zero-mean bivariate normal distribution
- estimate cutoffs, sigma, and correlation coefficient \( r \)

Aside: The Tetrachoric Correlation Coefficient

- The debate:
  - Yule: assuming underlying continuous normal variables is bogus
  - Pearson:
    - If there is actually a bivariate normal $Q \neq R$ (depending on cutoffs)
    - $Q$ is not unique
Aside: The Tetrachoric Correlation Coefficient

- No obvious reason to favour one approach, why do they differ?
- Pearson was a social Darwinist, committed to eugenics
  - Regression was created to measure heritability
  - The measure of correlation must be such that the effects of natural (or unnatural) selection can be predicted

  “if the theory of correlation can be extended [to categorical characteristics] we shall have much widened the field within which we can make numerical investigations into the intensity of heredity” — Pearson

Mathematical contributions to the theory of evolution.—VII. On the correlation of characters not quantitatively measurable," Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 195, no. 262-273 (1900): 1-47.
Aside: The Tetrachoric Correlation Coefficient

The choice of measures — even those as basic as a correlation coefficient — can be motivated by concerns and have effects which are profoundly ethical.
Example 2: Gender Bias in Word Embeddings

- Word embeddings represent words as vectors derived from their co-occurrence matrix (e.g. word2vec, later GloVE)
- Similar words have similar vectors, we can do algebra with vectors:
  - e.g. King - Man + Woman = Queen

Example 2: Gender Bias in Word Embeddings

- Word embeddings represent words as vectors derived from their co-occurrence matrix (e.g.
  word2vec, later GloVE)
- Similar words have similar vectors, we can do algebra with vectors:
  - e.g. King - Man + Woman = Queen
  - More specifically, 3CosAdd: $\arg\max_{a} (\cos(d, c) - \cos(d, a) + \cos(d, b))$

Example 2: Gender Bias in Word Embeddings

- Generate analogies for he::she, get crowdsourced workers to rank how stereotypical they are:
  - Examples: surgeon::nurse, Karate::Gymnastics, carpentry::sewing
- Suggestions to reduce debias of already-trained embeddings

Example 2: Gender Bias in Word Embeddings

But:

- 3CosAdd is broken
- For analogy A : B :: C : D word2vec implementation does not return D=B
  - This also applies to Bolukbasi’s direction-based formulation
- People choose which analogies to report: Manzini et al. found biased examples even with a mistakenly reversed the query (e.g. caucasian is to criminal as black is to X)

Nissim, Malvina, Rik van Noord, and Rob van der Goot. "Fair is better than sensational: Man is to doctor as woman is to doctor." arXiv preprint arXiv:1905.09866 (2019).
Example 2: Gender Bias in Word Embeddings

- A mixed conclusion
  - Of course there is gender bias in society
  - And there’s probably bias of some sort in word embeddings
  - But analogy tasks aren’t the right task to capture them
- More than that, analogy tasks are tricky to use for evaluation in algorithms

What should we do about these problems?
Can I have a checklist?

a) No

b) You can have some abstractions (but know that they are leaky)
Overview: Ethical Frameworks

- Research Ethics:
  - e.g. Belmont Report, Menlo Report
- Business Ethics
- Technology Ethics
- Engineering Ethics
  - e.g. AAAI
HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)

SITUATION:
THERE ARE
14 COMPETING
STANDARDS.

14?! RIDICULOUS!
WE NEED TO DEVELOP
ONE UNIVERSAL STANDARD
THAT COVERS EVERYONE’S
USE CASES. YEAH!

SOON:

SITUATION:
THERE ARE
15 COMPETING
STANDARDS.

https://xkcd.com/927/ CC-BY-NC
The Markkula Center *Framework for Ethical Decision-Making*

Central insights:

- There are different ethical perspectives
- Process matters: do the moral math

https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/a-framework-for-ethical-decision-making/
The Markkula Center *Framework for Ethical Decision-Making*

1. Recognize an Ethical Issue  
   a. Could it harm people?  
   b. Is it about ethics (contra law, efficiency, aesthetics, etc.)?  
2. Get the facts  
   a. Is there enough information?  
   b. Stakeholders  
   c. Options  
3. Evaluate options following different approaches  
   a. Deontology, Consequentialism, Virtue Ethics, Confucian, Buddhist, Hindu ethics  
4. Make a decision and test it  
   a. The Moral Math: explain how the decision is derived from the facts and evaluations  
5. Act and reflect on the outcome  
   a. Did it work?  
   b. What did we learn

https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/a-framework-for-ethical-decision-making/
The Markkula Center Approaches

Approaches are be academic philosophical schools, but more broadly different perspectives which focus on different aspects of the problem

- Deontology: e.g. Kant’s Categorical Imperative
- Consequentialism: e.g. Utilitarianism
- Virtue Ethics: character and habits
- Non-Western frameworks: e.g. Confucian, Buddhist, Hindu
Doing the ‘Moral Math’
That’s kind of vague...

a) Yes
b) Necessarily so
The Most General Advice: Reflective Equilibrium

Moral Principles → Moral Judgments → Moral Theories

See https://plato.stanford.edu/entries/reflective-equilibrium/
Resources

- https://datascience.columbia.edu/FATES-Elaborated
- https://www.fatml.org/resources
- https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/a-framework-for-ethical-decision-making/
Current Consumer Privacy Legislation

GDPR (General Data Protection Regulation)

CCPA (California Consumer Privacy Act)

In 2019 Many other there have been bills introduced or filled in at least 25 states, as we continue in this new privacy paradigm, Consumers will have more control and access to their personal information and how tech companies can use it.

What does this mean for me?