1 □ **Can a Machine Be**
Racist or Sexist?

On Social Bias in Machine Learning

Renée M. P. Teate

HelioCampus

2 □

1. This talk: Introducing concepts and examples of social bias in machine learning to get us all on the same page
   (~30 mins)
2. Panel Discussion w/ Audience Q&A (~20 mins)

3 □ **What comes to mind**
for most people
when they are asked about
their fears related to
“Artificial Intelligence”
or “Machine Learning”?

4 □

5 □ **So what is *already* going on**

with AI and Machine Learning
that *should* concern us?

6 □ **And are the impacted people and communities aware of what’s already happening?**

Are the people who design these systems aware of the possible impacts of their work
on people’s lives as they
design and deploy data products?

7 □ **“But if we take humans out of the loop and leave decisions up to computers, won’t it reduce the problems inherent in human decision-making?”**

8 □ **Can a Machine Be**
Racist or Sexist?

9 □ **Can a Machine Learning Model**
Be *Trained* to Be

Racist or Sexist
(or made biased or injust
in other ways --
***intentionally or not*)?**

10 □ **Let’s define**

11 □ **Machine [Algorithm]**

“a step-by-step procedure for solving a problem”

Merriam-Webster Dictionary

12 □ **Racism**

“racial prejudice or discrimination”

“a belief that race is the primary determinant of human traits and capacities and that racial differences produce an inherent superiority of a particular race”

Merriam-Webster Dictionary

13 □ **Racism**

“a doctrine or political program based on the assumption of racism and *designed to execute its principles*”

[or, not designed NOT to execute its principles!]

14 □ **Sexism**

“prejudice or discrimination based on sex”

“behavior, conditions, or attitudes that foster stereotypes of social roles based on sex”

Merriam-Webster Dictionary

15 **Institutional or Systemic Racism and Sexism**

a system that codifies and perpetuates discrimination
against individuals or communities
based on their race or sex

(note: these systems are designed/engineered by people)

16 **Statuses Protected by Laws in the U.S.**

1
- Race
- Sex
- Religion
- National Origin
- Age
- Disability Status

2
- Pregnancy
- Citizenship
- Familial Status
- Veteran Status
- Genetic Information

17 **Example: Bank Loans Before Machine Learning**

Bank officer deciding whether to give a loan, assessing likelihood to repay:

- Employment Status and History
- Amount of Debt and Payment History
- Income & Assets
- “Personal Character”
- Co-Signer
- References
- Credit Score
    - Based on amount of debt, credit card payment history, debt-credit ratio etc.
    - May seem fair, but remember things like on-time payment of rent not included
    - Feedback loop - no/bad credit history, can’t get credit, high interest, can’t improve credit score
    - Is somewhat transparent, and errors can be corrected

18 **Example: Bank Loans With Machine Learning**

1 *Algorithm* assessing likelihood to repay:

2
- Employment Status and History
- Amount of Debt and Payment History
- Income & Assets
- ~~"Personal Character"~~
- Co-Signer
- ~~References~~
- Credit Score
- Detailed Spending Habits
  - Expenditures per month
  - Where you shop
  - Bill payment patterns

3
- Where You Live
- Social Media Usage
- Time You Wake Up
- Workout Consistency
- Driving Habits
- Time Spent Playing Video Games
- Favorite Music
- Browser History
- Facebook Friends' Financial Status
- etc etc etc

19 □ **Is it fair for your interest rate,**
or whether you even get a loan offer,
to be based on the default rates of "similar" people who, for instance, listen to the same kind of music as you?

20 □ **What does it mean for decisions to become increasingly data-driven and automated?**

We're still making the same types of decisions
(Who should receive funds from government programs? Who is at risk of dropping out of a university, and how do we intervene? What medical treatment should be applied based on a set of symptoms? Where should we locate the next branch of our business? etc etc),
but now we're using much more data,
and programming computers to help us find patterns
from datasets larger than humans could sensibly process.

21 □

If designed well,

machine learning systems
can improve our world!

Better more targeted answers faster!

More efficient use of taxpayer dollars, students receiving financial aid and intervention tutoring to help keep them in school, highly customized medical treatments, lower-risk business decisions!

22

But we have to keep in mind that now:

“...we have the potential to make bad decisions far more quickly, efficiently, and with far greater impact than we did in the past”

-Susan Etlinger, 2014 TED Talk

23 **How can human biases get into a machine learning model?**

Let’s explore how machine learning systems are designed and developed

24 **Some Types of**
Machine Learning Models

25 **Regression**

What output value do we expect
an input value to translate to?

26

27

28

29

30

31

Forecasting Record-Breaking Long Jump Distance by Year

“Olympics Physics: The Long Jump and Linear Regression”

https://www.wired.com/2012/08/physics-long-jump-linear-regression/

32

Predicting Natural Disaster Damage by Counting Relevant Social Media Posts

“Rapid assessment of disaster damage using social media activity”

http://advances.sciencemag.org/content/2/3/e1500779/tab-figures-data

33 □ **Classification**

Which group does X belong to?

34 □

35 □

36 □

37 □

38 □

39 □

40 □

41 □

42 □

43 □

44 □

Is there an animal in this camera trap image?

“Deep learning tells giraffes from gazelles in the Serengeti”

https://www.newscientist.com/article/2127541-deep-learning-tells-giraffes-from-gazelles-in-the-serengeti/

45 □

Is a crime scene gang-related?

“Artificial intelligence could identify gang crimes—and ignite an ethical firestorm”

http://www.sciencemag.org/news/2018/02/artificial-intelligence-could-identify-gang-crimes-and-ignite-ethical-firestorm

46 □ **Clustering**

How should we group this data?

47 □

48

49

50

51

52

How might we segment our customers?

“Understanding, Analyzing, and Retrieving Knowledge from Social Media”

http://cucis.ece.northwestern.edu/projects/Social/

53

Which neighborhoods are most likely to experience home burglaries this month?

“Crime Forecasting Using Spatio-Temporal
Pattern with Ensemble Learning”

https://www.cs.umb.edu/~csyu/YU_resume%202016_01_08_files/yuPAKDD2014.pdf

54 **Artificial Neural Networks**

Reinforcement Learning

Collaborative Filtering

etc.

55 **The purpose of most of these algorithms is to find patterns, trends, group things that are similar…**

In other words, we’re basically asking the computer to use lots of information to make generalizations, or stereotype.

56

57

58

59 **Predictive Model Development Process**

60 **Predictive Model Development**

- Deciding what you’re predicting / optimizing for
- Data collection and storage
- Data cleansing/preparation

- Feature selection & engineering
- Importing data into different algorithmic models
- Training & Testing
- Model evaluation & competition; Deciding what qualifies as a "good model"
  - Parameter Tuning, Cost function, Selecting cutoff values or stopping conditions, etc
- "Productionizing" - Applying to live data, building interactive reports for end-users, explaining what the results mean and how to use them to make decisions
- Monitoring, Improving, and Re-training over time

61 □

62 □ **Where within this process can social biases be introduced?**

63 □ **Data Collection: Incorrectly Recorded**

64 □ **Data Collection: Manipulated**

65 □ **Data Collection: Not Representative**

66 □ **Update...**

67 □ **Data Collection: Contains Historic Biases**

68 □ **Data Availability: Imbalanced Dataset**

69 □ **Model Evaluation: Confusion Matrix & Cost**

70 □

71 □ **Model Evaluation: Confusion Matrix & Cost**

72 □ **Data Pre-Processing: Dropping Data**

73 □ **Model Training: Bias Amplification**

http://vicenteordonez.com/files/bias.pdf

74 □ **Feature & Algorithm Selection - Different algorithms handle different types of data in different ways**

Target Selection/Optimization Goal - What are you optimizing for? (Technical & Business Decision)

Model Evaluation - How good does your model have to be to decide to stop improving it? And how do you define "good"?
Consider "cost" of each type of error.

75 □ **Example: Optimizing for maximum video viewing time may incentivize the display of alarming/intriguing information, whether or not it is true (propaganda)**

76 □ **Implementing the Trained Model - In what scenarios can your model be applied? How generalizable is it?**

Interpretation - How do you interpret the results? How do you document and explain to others how to interpret the results?

Maintenance - For how long can the current model be applied? When does the model need to be retrained? Does the "ground truth" change?

77 □ **Can your model be gamed?**

78 □ **Could your model cause harm?**

Or perpetuate existing social hierarchies,
preventing a fair playing field?

79 □ **Some Types of Harm a Model Can Perpetuate**

Allocative harms - resources are allocated unfairly or witheld (transactional, quantifiable)

Representational harms - systems reinforce subordination/perceived inferiority of some groups (cultural, diffuse, can lead to other types of harm)

- stereotyping
- underrepresentation
- recognition
- denigration
- Ex-nomination

(from Kate Crawford's talk at 2017 NIPS Conference, The Trouble With Bias)

80 □

81 □ **What makes a model a "Weapon of Math Destruction"?**

- Opacity - inscrutable "black boxes" (often by design)

- Scale - capable of exponentially increasing the number of people impacted

"The privileged...are processed more by people,
the masses by machines."

- Damage - can ruin people's lives and livelihoods

82 □ **So, Can a Machine Be**
Racist or Sexist?

83 □ **YES**

84 □

85 □ **What can we do about it?**

86 □

- Be aware of the potential for bias and disparate impact machine learning models can perpetuate (and help educate others)
- Test for bias in your models & evaluate model performance on minority classes in your dataset
- Evaluate possible uses (or misuses) of your model, and "perverse incentives" it may create in the system into which it's being deployed
- Improve transparency & explainability of how your model categorizes and predicts things (LIME)

87 □

- Document Data Source & Transformation Pipeline to help manage data governance & provenance and allow reproducible research
- Communicate context, and explain model generalization limits to end-users
- Involve Domain Experts who know the history of the data collection, actual field definitions, data quality issues, system changes over time, how the model will likely be applied, ethical issues and laws in the field of application, etc.
- Gather Representative Training Data

88 □

- Include fairness as an optimization objective, add social bias penalties (research emerging, more needed)
- Research and Develop new tools and techniques for detecting bias and reducing disparate impact caused by machine learning models
- Build adversarial tools to stress-test your own models, and thwart others that are causing harm
- Hire Diverse Teams
- Demand Accountability and Regulation in the industry, and in your own organizations

89 □ **Organizations working on this**

AI Now Institute at NYU https://ainowinstitute.org/

Algorithmic Justice League https://www.ajlunited.org/
(Watch Joy Buolamwini's TED Talk!)
Data for Democracy, Bloomberg, BrightHive
Data Science Code of Ethics
http://datafordemocracy.org/projects/ethics.html
Fairness, Accountability, and Transparency in Machine Learning https://www.fatml.org/

90 **Example research**

91 **More resources**

Flipboard Magazine where I collect articles on this topic:
https://flipboard.com/@becomingdatasci/bias-in-machine-learning-rv7p7r9ry

https://www.becomingadatascientist.com/2015/11/22/a-challenge-to-data-scientists/
https://developers.google.com/machine-learning/fairness-overview/
https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/
https://www.fatml.org/resources/relevant-scholarship
https://twitter.com/random_walker/status/961332883343446017

92

"How do we govern ourselves? How do we instill that trust in others that we as stewards of that data understand the power that we have, and want to make sure that we're doing right by the people who are trusting us with that data?"
-Natalie Evans Harris
Data for Good Exchange 2017, Inside Bloomberg

93

Renée Teate
BecomingADataScientist.com
@becomingdatasci

94 **Social Bias in Machine Learning Panel**