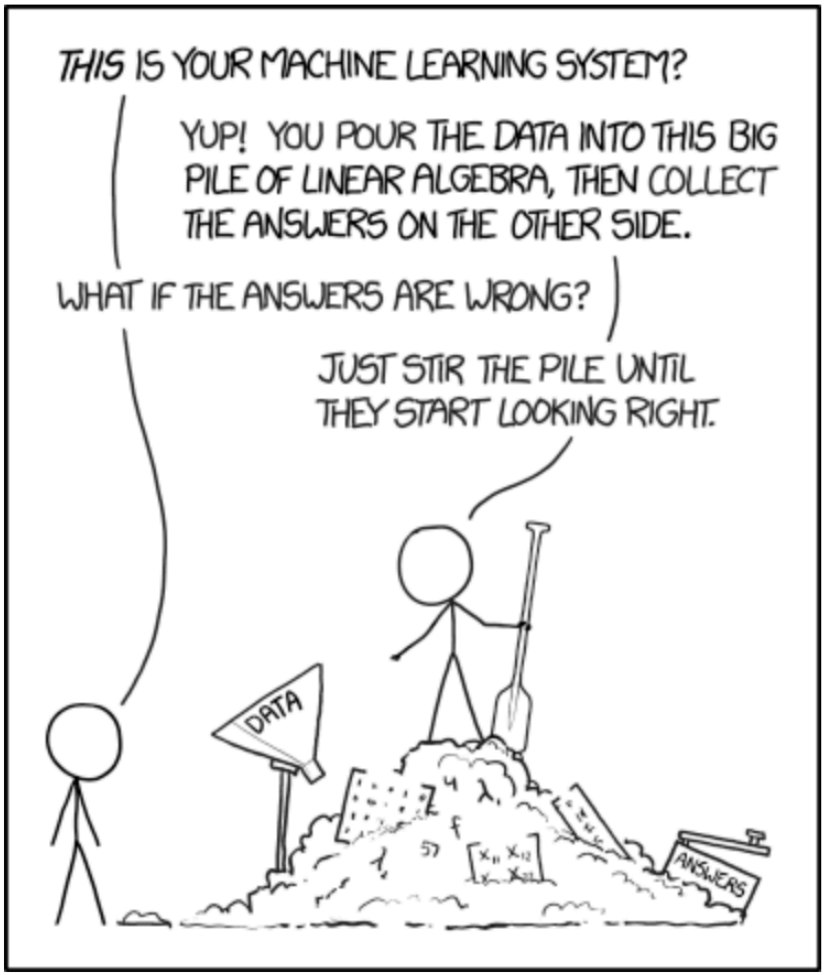
Ethics and AI

Yulia Newton, Ph.D.

CS156, Introduction to Artificial Intelligence

San Jose State University

Spring 2021



https://xkcd.com/1838

What do we want from AI?

- European Union High Level Expert Group on Artificial Intelligence guidelines for model building:
  - AI models should be
    - Lawful - respecting all applicable laws and regulations
    - Ethical - reflect ethical principals and values
    - Robust - from technical and social environment perspectives

Ethics in AI

- “The ethics of artificial intelligence is the branch of the ethics of technology specific to artificially intelligent systems. It is sometimes divided into a concern with the moral behavior of humans as they design, make, use and treat artificially intelligent systems, and a concern with the behavior of machines, in machine ethics. It also includes the issue of a possible singularity due to superintelligent AI.” - Wikipedia

- When we talk about ethics applied to the field of AI we seek to promote discussions about ethical, regulatory, and policy implications that arise from the development of AI technologies

Motivational example of “unethical” AI

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is one of the best known examples that promotes ethical considerations when designing an AI system
  - A case management and decision support tool used by U.S. courts to assess the likelihood of a defendant becoming a recidivist
  - Was used by the states of New York, Wisconsin, California, Florida's Broward County, and some other jurisdictions

Motivational example of “unethical” AI (cont’d)

- Turns out this AI system was no better at predicting future crimes than asking random people (“The accuracy, fairness, and limits of predicting recidivism” Julia Dressel & Hany Farid, *Science Advances*)
  - This complicated deep AI system also did not better than very simple ML algorithms
- The algorithm turned out to be biased against African American individuals
  - African American offenders were twice more likely to be labeled as high risk than white offenders, who were likely to be labeled low risk

Motivational example of “unethical” AI (cont’d)

- So what went wrong?
- The algorithm itself was not actually wrong, it did not make mistakes
- Multiple research teams analyzed this issue and it was suggested that maybe there is simply no signal in the data
  - Maybe you simply cannot predict recidivism with high confidence from the data we have (age, gender, ethnicity, prior record, etc.)
  - Risk assessment tools add no value to an expert decision

Goals: AI should be …

- Diverse
- Inclusive
- Equitable

The ethics of predictions

- When we talk about AI being “ethical” we usually mean the predictions of a given AI model achieve the three goals we just talked about
  - Diversity, inclusivity, equity

Watch out for biases in the training data

- You can design a great AI model architecture but if your input data is biased then your model and its predictions will also be biased
  - E.g. class imbalances
  - Representative sampling
  - Your input data should be a representative sample of the population data
    - If you only include basketball players in your sample and height is an independent variable, then predictions for non-basketball players will be biased

Another example of ethical issues in AI (Allegheny Family Screening Tool)

- Allegheny Family Screening Tool - a system designed to assist experts in decision whether to remove a child from a family because of abusive circumstances
- Problem with the system:
  - Assessment to remove the child occurred 3 times more often for African American and bi-racial families than for white families, especially middle and upper class families
- Why did the problem occur?
  - Public dataset used to train this AI system contained biases
  - Middle and upper class families are better at hiding the abuse and therefore were not represented by the training dataset

Another example: Amazon's recruiting tool

- Amazon's AI-based recruiting tool, started in 2014, was aimed at automating the task of recruiting
  - Review the applicant's resume and rank this applicant
  - Eliminate manual sorting through the applications
- Problem with the system:
  - The system exhibited a bias against women
- Why did the problem occur?
  - Amazon used a historical data for the previous 10 years to train the AI model
  - The historical data contained biases against women in technology
- Amazon stopped using this tool in 2018
  - https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Algorithms have limits

- No matter how sophisticated and well-designed the AI algorithm is, it is only as good as the training data
- Biased data leads to biased AI models

Humans are often the source of the bias

- Humans design AI systems, therefore these systems reflect human biases
  - This is why team diversity is important on these types of projects
- Even the best AI system still contains some biases
  - The problem arises when these biases affect and bias predictions
- Type I vs. Type II errors
  - False positive vs. false negative errors and biases
  - E.g. is your breast cancer prediction model biased towards positive predictions?

Another example: healthcare risk assessment

- Widely used risk assessment algorithm for computer assisted health condition scoring showed a racial bias against African Americans
  - Designed to predict which patients need extra medical care
  - “Dissecting racial bias in an algorithm used to manage the health of populations” Obermeyer et al. 2019 Science
    - https://science.sciencemag.org/content/366/6464/447
  - At a given risk African American patients were considerably sicker than white patients
- What went wrong?
  - The algorithm designers used healthcare costs as a proxy for medical needs as the dependent variable in their training data
  - The algorithm was actually predicting healthcare costs rather than an illness
  - Based on historical data, less money was spent on caring for African American patients, therefore this variable did not correlate with the level of sickness for this group of patients

Another example: Facebook ads

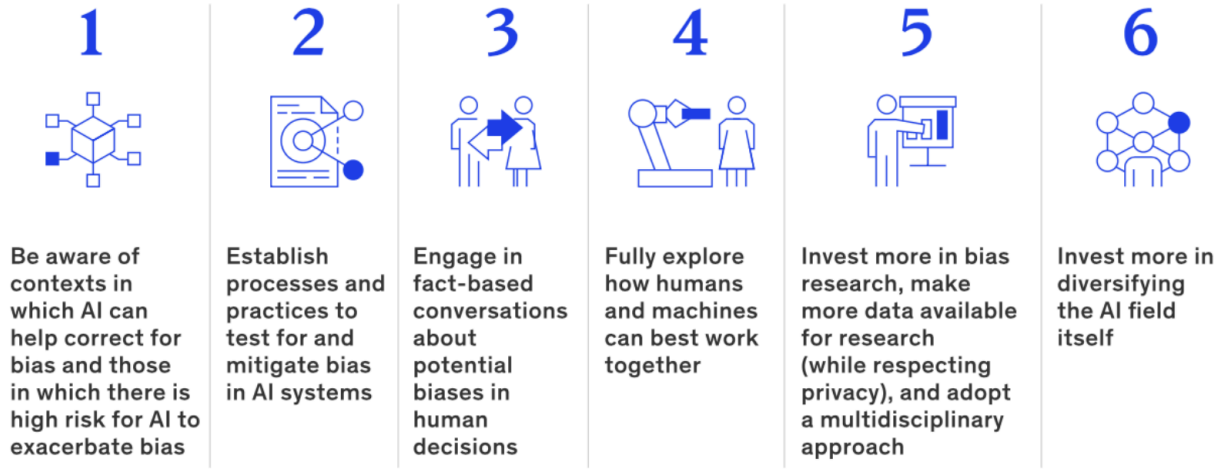
- In 2019 Facebook ads allowed companies to specify gender, race, and religion for targeting their advertisements to
- This lead to such trends as:
  - Women were prioritized for nursing and secretarial job ads
  - Men, particularly some minority groups, were prioritized for janitorial and taxi driving job ads
- Starting in 2020 Facebook no longer allows specifying these parameters for ad targeting
  - https://www.technologyreview.com/2019/03/20/1225/facebook-is-going-to-stop-letting-advertisers-target-by-race-gender-or-age

So, how do we move forward?

- We do our best to reduce biases
  - In training data
    - Feature selection
    - Feature engineering
    - Proxy variables
  - In human design of
    - Algorithms
    - Hyperparameters

Some things to consider

Six potential ways forward for artificial-intelligence (AI) practitioners and business and policy leaders to consider



McKinsey & Company

SOURCE: MCKINSEY

Some tools to assess biases in models/data

- IBM's open source library **AI Fairness 360**
  - https://github.com/Trusted-AI/AIF360
- IBM's **Watson OpenScale**
  - https://www.ibm.com/cloud/watson-openscale/model-risk-management
- Google's What-if tool
  - https://pair-code.github.io/what-if-tool/index.html

Additional resources

- Krita Sharma’s Ted Talk
  - https://youtu.be/BRRNeBKwvNM
- Barak Turovsky’s talk
  - https://youtu.be/gU2VgMfciQA

Other ethical considerations when it comes to AI

- Job loss and wealth inequality
- Blind trust in AI systems
- How much freedom and responsibilities can we afford to an AI system?
  - Almost perfect performance may still lead to large numbers of mispredictions
- What if AI goes rouge?
- How should we treat AI? What rights does AI have?
-

Conclusion

- As an AI and ML practitioner, it is your job to constantly be aware of and question biases and their sources in your AI system/model
- Use steps to mitigate those potential biases upfront