ML Ethics, Final Summary

Jindřich Libovický

January 9, 2024

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Today's Lecture Objectives

After this lecture you should be able to

- Explain what main theoretical ethical frameworks are.
- Reason about ethical problems in various stages of developing ML System.

Why ML Ethics? (1)

ML makes things faster, cheaper, more efficient, more accurate, more accessible,

which seems morally good or neutral...

ComputerHistory.org: Thomas J. Watson meets Adolf Hitler

...but not always is.

(IBM computers helped make administration and logistics of the Nazi Germany more efficient.)

Why ML Ethics? (2)

- There are high-stake applications of ML: decision making (in public and business administration)

- People tend to consider algorithm outputs as objective

- There are bad actors (military use, massive surveillance, disinformation, ...)

- Inherent problems of ML: Good intentions might lead to unintended harm (biases, lack of explainability)

What is Ethics

Ethics or moral philosophy is a branch of philosophy that "involves systematizing, defending, and recommending concepts of right and wrong behavior".

- Simply taken: Study of what is **right and wrong**
- Several major theoretical frameworks:
  - **Deontological** ethics – rule/principle driven, good action follows rules
  - **Consequentialist** ethics / utilitarism – consequences matter, good action has good consequences
  - **Virtue** ethics (encourage curiosity, creativity, solidarity, etc.) – good action is what a virtuous person would do
  - **Contract** ethics (everything is social contract) – good action is what people implicitly agreed upon

Deontological Ethics

- Focuses on the **inherent nature of actions** rather than their consequences
- Involves adhering to predefined **rules and principles** (e.g, the Universal Declaration of Human Right, the Ten Commandments, Kant's Categorical Imperative)
- In the ML context, it typically means principles like: beneficence, non-malevolence, privacy, non-discrimination, autonomy + informed consent

Pros

1. Clear guidelines
2. Stability
3. Respect for rights

Cons

1. Rigidity
2. Conflict of principles
3. Neglect of consequences

- Criticism: When you are evil, you can always argue that you build on good principles regardless of the consequences.

Utilitarian Ethics

- An ethical theory that emphasizes the maximization of **overall happiness or well-being** or minimizing harm
- It focuses on **consequences of actions**
- Good decisions = decisions that lead to the greatest overall positive impact.

Pros

1. Flexibility
2. Quantifiable

Cons

1. Overlooking individual rights
2. Difficulty in defining the objective

- Implicitly behind most work on ML ethics that focuses on harmful consequences for various social groups
- Gets tricky once we consider very low-probability events with very high impact

Different Theories, Different Ideas

Examples of different ways of thinking under different theoretical frameworks:

Deontological ethics

Never use ethnicity as a feature.

Build trust.

Insist on privacy, fairness and justice.

Utilitarian Ethics

Who will be effected and how?

What harm can happen: psychological, political, environmental, moral, cognitive, emotional...

How to prevent the harm?

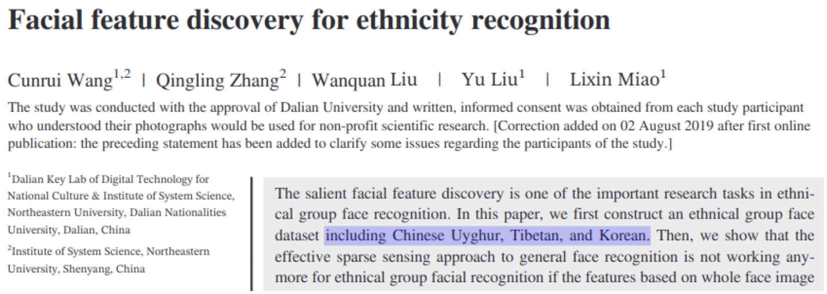
Stages of ML development

Ethical problems might emerge in all stages in ML system development

- Problem definition – some tasks are inherently problematic
- Data collection – biases in data, unethical collection
- Model development
- Model evaluation – metrics do not cover important things
- Model deployment – use outside of original scope, feedback loops

Problem Definition

Sometimes is a bad idea to use ML in a first place.



Facial feature discovery for ethnicity recognition

Cunrui Wang[1,2] | Qingling Zhang[2] | Wanquan Liu | Yu Liu[1] | Lixin Miao[1]

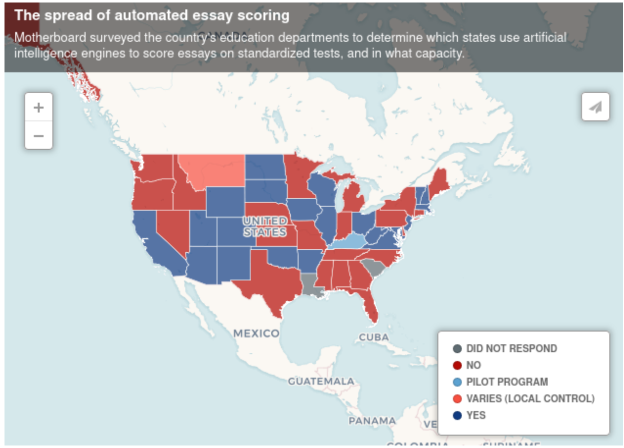
The study was conducted with the approval of Dalian University and written, informed consent was obtained from each study participant who understood their photographs would be used for non-profit scientific research. [Correction added on 02 August 2019 after first online publication: the preceding statement has been added to clarify some issues regarding the participants of the study.]

[1]Dalian Key Lab of Digital Technology for National Culture & Institute of System Science, Northeastern University, Dalian Nationalities University, Dalian, China

[2]Institute of System Science, Northeastern University, Shenyang, China

The salient facial feature discovery is one of the important research tasks in ethnical group face recognition. In this paper, we first construct an ethnical group face dataset including Chinese Uyghur, Tibetan, and Korean. Then, we show that the effective sparse sensing approach to general face recognition is not working anymore for ethnical group facial recognition if the features based on whole face image

Many schools in the US use automatic essay scoring for Graduate Record Examinations (GRE).

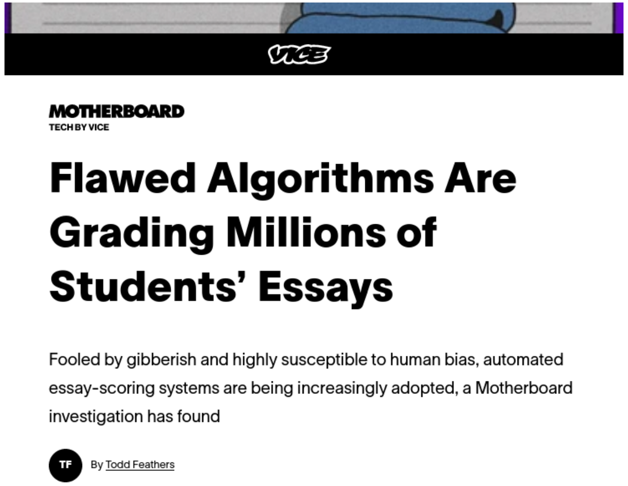


The spread of automated essay scoring

Motherboard surveyed the country's education departments to determine which states use artificial intelligence engines to score essays on standardized tests, and in what capacity.

Lack of transparency: students have right to know why they were accepted

Allows **metric gaming** if you guess what the features might be



MOTHERBOARD
TECH BY VICE

Flawed Algorithms Are Grading Millions of Students' Essays

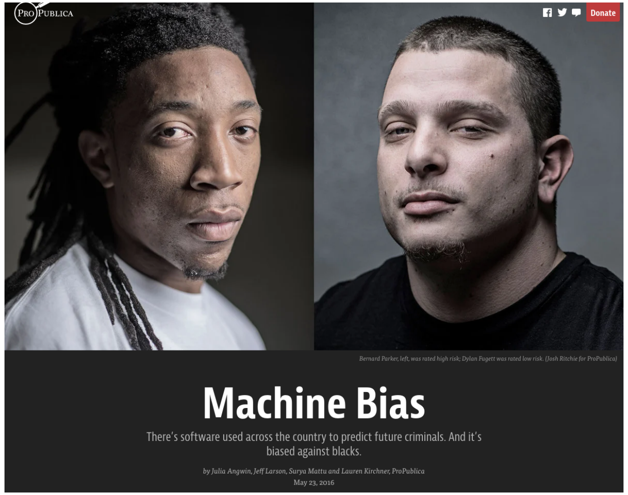
Fooled by gibberish and highly susceptible to human bias, automated essay-scoring systems are being increasingly adopted, a Motherboard investigation has found

TF By Todd Feathers

https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays

Recidivism prediction: COMPAS

- Proprietary ML-based software using 130 unknown features is used in several US states to predict recividism of individual defendants.
- Extrinsic evaluations show that similar cases with defendants of different enthnicity then to favor white defendants.
- A simple model considering previous criminal record, age and education level performs similarly, as well as judgment of non-professional individuals.



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

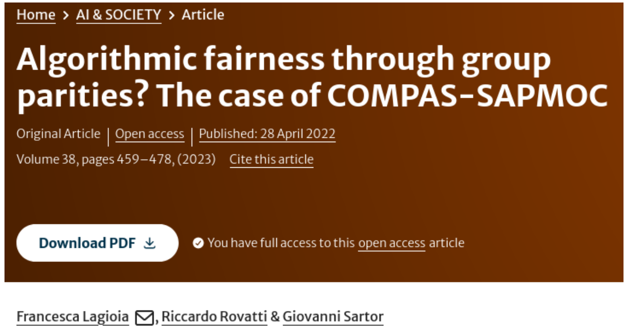
Machine Bias

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

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



Home > AI & SOCIETY > Article

Algorithmic fairness through group parities? The case of COMPAS-SAPMOC

Original Article | Open access | Published: 28 April 2022

Volume 38, pages 459–478, (2023) Cite this article

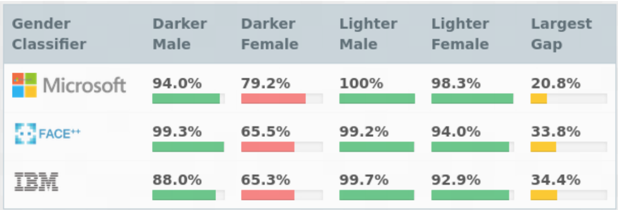
Download PDF
You have full access to this open access article

Francesca Lagioia, Riccardo Rovatti & Giovanni Sartor

https://link.springer.com/article/10.1007/s00146-022-01441-y

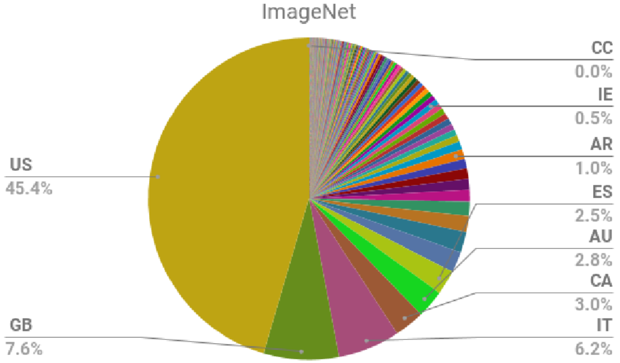
Data Collection & Biases in Data

- Representation bias: The data might not be representative of the population (missing minorities, poor people, …)
- Data (and especially text) from the Internet does not represent the world as it is (only those who have access and are loud) and the world as it should be
- Historical bias: Inequalities from the past when the data was created are preserved in the datasets
- Copyright issues, especially with generative models

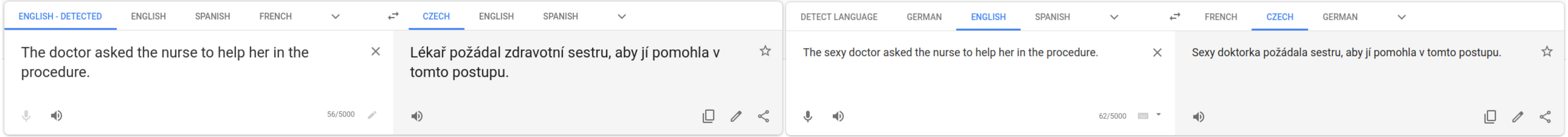


Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on fairness, accountability and transparency. PMLR, 2018.



https://arxiv.org/pdf/1711.08536.pdf



The doctor asked the nurse to help her in the procedure.

Lékař požádal zdravotní sestru, aby jí pomohla v tomto postupu.

The sexy doctor asked the nurse to help her in the procedure.

Sexy doktorka požádala sestru, aby jí pomohla v tomto postupu.

Libovický, Jindřich. "Neuronové sítě a automatický překlad." Rozhledy matematicko-fyzikální 94.4 (2019): 30-40.

Problematic Practices by Data Collection

Crowdsourcing

- People are hired to do the job the ML model will do
- Not well payed (often in third world countries), monotonous work, occasionally causing psychological harm
- Gig economy: what was originally meant as earning extra cash becomes full-time job without labor protection

Crawford, Kate. The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press, 2021. Chapter 2.

Log mining and user data collection

- Training data is collected from users that have no other choice than provide data by using services (not using them and keeping social/work/political life at the same is impossible)
- Nontransparent transaction: user gets service (for free or payed) and provides data

Couldry, Nick, and Ulises A. Mejias. The costs of connection: How data is colonizing human life and appropriating it for capitalism. Stanford University Press, 2020.

Model Development

- Discretization of outputs might lead to bias amplification

  Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (Zhao et al., EMNLP 2017)

- Larger models are more prone to overfitting: might lead to memorization of very specific patterns (privacy issues, remembering particular names)

  Are Large Pre-Trained Language Models Leaking Your Personal Information? (Huang et al., Findings EMNLP 2022)

- Distilled models are prone to stereotyping

  Why Knowledge Distillation Amplifies Gender Bias and How to Mitigate from the Perspective of DistilBERT (Ahn et al., GeBNLP 2022)

- Models might learn protected attributes by proxies (e.g., ethnicity from name, school)

  Adversarial Removal of Demographic Attributes Revisited (Barrett et al., EMNLP-IJCNLP 2019)

Model Evaluation

- Metrics might not capture everything we need
  - E.g., translation fluency does not capture gender bias
  - Macro-averaging might hide bad performance for specific user groups (typically minorities)
- Human Resources: employment recommendation based on CV
  - Precision: The business implies optimizing for precision – you only recommend few candidates and they need to be the good ones.
  - No one sees the recall – which would show that models might discriminate against gender, age, ethnicity, etc.

Proxy metrics optimizing something else

Paperclip Maximizer: a 2003 thought experiment by Nick Bostrom.

Wrongly chosen objective for an AI-driven paperclip factory that wipes out humanity and turns the entire Solar system into a paper-clip factory.

- Platforms like YouTube use watch time as a proxy for content quality (and btw. more watch time brings them more money)
- Non-profit algotransparency.org monitors stats on YouTube recommendations: 2016-2018 most recommended videos supporting alternative narrative on political events (US and French elections, mass shootings)
- Presumably, this was the type of content maximizing the watch time

https://guillaumechaslot.medium.com/how-algorithms-can-learn-to-discredit-the-media-d1360157c4fa

Use of model in practice & Feedback loops

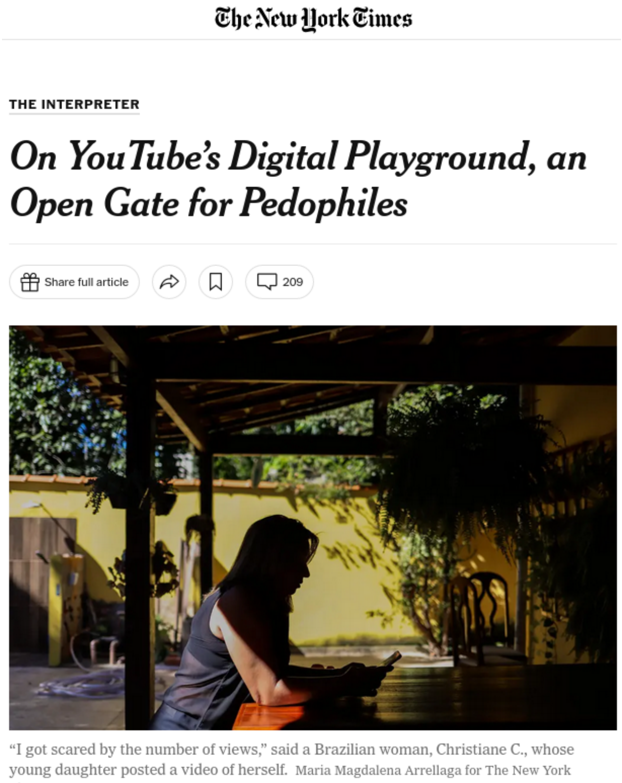
Train and test data mismatch

Minority language is more often classified as hate speech/NSFW.

The Risk of Racial Bias in Hate Speech Detection (Sap et al., ACL 2019)

Feedback loops

- Recommender systems: Predictions determine future user behavior that are used as training data again.
- Leads to very specific echo chambers and self-affirmative groups.
- NY Times: YouTube's recommendation discovered a "category" of home videos of barely clothed children sought out by pedophiles.



The New York Times

THE INTERPRETER

On YouTube's Digital Playground, an Open Gate for Pedophiles

"I got scared by the number of views," said a Brazilian woman, Christiane C., whose young daughter posted a video of herself. Maria Magdalena Arrellaga for The New York

https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html

Today's Lecture Objectives

After this lecture you should be able to

- Explain what main theoretical ethical frameworks are.
- Reason about ethical problems in various stages of developing ML System.

Theoretical concepts

Basic statistic

Bernoulli distribution, Categorical distribution, Normal Distribution, descriptive statistics (mean, variance, correlation), Maximum Likelihood Estimation, Bayes Theorem

Information theory basics

Entropy, Conditional Entropy, Cross-Entropy, KL-Divergence, Mutual Information

Optimization

Set derivative to zero, Lagrange multipliers, numerical optimization with SGD and second-order methods

Machine learning methodology

Working with data

- **Data annotation**: inter-annotator agreement (correlation, Cohen's alpha)
- **Features**: numerical/categorical features, polynomial features, TF-IDF, use pre-trained embeddings, representation learning
- **Normalization**: min-max scaling, standardization, whitening

Training and Evaluation

- **Data splits**: optimization for unseen data, train, validation, test split
- **Overfitting and regularization**: early stopping $+$ reading learning curves, L^2-regularization, dropout in MLP, prior in Bayesian models
- **Evaluation metrics**: accuracy, mean squared error, precision/recall/F-score, correlation, hypotheses testing

Machine learning models

Geometric intuition

Linear regression, Perceptron, Nearest Neighbors Classification and Regression, SVD, k-Means clustering

Probabilistic intuition

Linear regression, logistic regression, Multi-layer Perceptron, Naive Bayes, PCA

Decision trees

Random forest, Gradient boosted decision trees

Course Objectives: What you hopefully learned

After this course you should...

- Be able to reason about task/problems **suitable for ML**
  - Know when to use classification, regression and clustering
  - Be able to choose from this method Linear and Logistic Regression, Multilayer Perceptron, Nearest Neighbors, Naive Bayes, Gradient Boosted Decision Trees, k-means clustering
- Think about learning as (mostly probabilistic) **optimization on training data**
  - Know how the ML methods learn including theoretical explanation
- Know how to properly **evaluate** ML
  - Think about generalization (and avoiding overfitting)
  - Be able to choose a suitable evaluation metric
  - Responsibly decide what model is better
- Be able to **implement ML algorithms** on a conceptual level
- Be able to **use Scikit-learn** to solve ML problems in Python