PURDUE UNIVERSITY | Department of Computer Science

CS37300
Data Mining & Machine Learning

Ethics Issues in Data Mining

Profs. Steve Hanneke and Chris Clifton

5 April 2023

Indiana Center for Database Systems ™

PURDUE UNIVERSITY
Department of Computer Science

Ethics Issues for Data Mining & ML
What's the Problem?

- Privacy
  - Training data
  - Allowed uses
- Fairness
  - Inequitable outcomes
  - Variance in accuracy
- Data inaccuracy
- Explainability
- Redress
  - What if someone disputes results?

23

PURDUE UNIVERSITY
Department of Computer Science

Discrimination in AI: What's all the fuss?


Facebook Thinks Some ... in Online Ad Delivery

Amazon scraps secret AI that showed ...

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

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Jörg Carstensen—AP

Some Native Americans say Facebook won't allow them to log in because their names are "inauthentic."

PURDUE UNIVERSITY
Department of Computer Science

What's all the fuss?

(Dastin '18)

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

- Resume screening tool
  - Trained on prior applications
  - Demonstrated bias toward male applicants
  - Manual avoidance of "obvious" discriminatory words
- *Scrapped for fear of remaining biases*

PURDUE UNIVERSITY
Department of Computer Science

What's all the fuss?
(Angwin, Larson, Mattu, Kirchner '16)

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

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

- Similar cases lead to different outcomes
  - Minor theft (shoplifting, stealing a bike)
  - Black offender predicted as more likely to commit future crime than white
  - *Despite white offender having criminal record!*
- Statistical analysis suggests this is common

PURDUE UNIVERSITY
Department of Computer Science

What's all the fuss?
(Sanburn '15)

Facebook Thinks Some Native American Names Are Inauthentic

Josh Sanburn @joshsanburn | Feb. 14, 2015

The social network is barring some Native Americans from logging in

If you're Native American, Facebook might think your name is fake.

The social network has a history of telling its users that the names they're attempting to use aren't real. Drag queens and overseas human rights activists, for example, have experienced error messages and problems logging in in the past.

Jörg Carstensen—AP

Some Native Americans say Facebook won't allow them to log in because their names are "inauthentic."

- Ms. Lone Elk (and others) required to provide identification to use Facebook
  - Viewed as potential violation of "real name" policy
- No such barriers for "dominant majority"

PURDUE UNIVERSITY
Department of Computer Science

What's all the fuss? (*Sweeney '13*)

Discrimination in Online Ad Delivery

Latanya Sweeney
Harvard University
latanya@fas.harvard.edu

January 28, 2013[1]

Abstract

A Google search for a person's name, such as *"Trevon Jones"*, may yield a personalized ad for public records about Trevon that may be neutral, such as *"Looking for Trevon Jones? ...",* or may be suggestive of an arrest record, such as *"Trevon Jones, Arrested?..."*. This writing investigates the delivery of these kinds of ads by Google AdSense using a sample of racially associated names and finds statistically significant discrimination in ad delivery based on searches of 2184

- Blacks and whites see different ads on the internet
  - *Even if race not part of the profile*
- Sweeney found that first names typically associated with blacks and whites lead to different ads
  - Otherwise identical profiles and histories

PURDUE UNIVERSITY
Department of Computer Science

What's all the fuss? (*Datta,Tschantz, and Datta '15*)

DE GRUYTER OPEN Proceedings on Privacy Enhancing Technologies 2015; 2015 (1):92–112

Amit Datta*, Michael Carl Tschantz, and Anupam Datta

Automated Experiments on Ad Privacy Settings

A Tale of Opacity, Choice, and Discrimination

Abstract: To partly address people's concerns over web tracking, Google has created the Ad Settings webpage to provide information about and some choice over the profiles Google creates on users. We present AdFisher, an automated tool that explores how user behaviors, Google's ads, and Ad Settings interact. AdFisher can run browser-based experiments and analyze data using machine learning and significance tests. Our tool uses a rigorous experimental design and statistical analysis to ensure the statistical soundness of our results. We use AdFisher to find that the Ad Settings was opaque about some features of a user's profile, that it does provide some choice on ads, and that these choices can lead to seemingly discriminatory ads. In particular, we found serious privacy concern. Colossal amounts of collected data are used, sold, and resold for serving targeted content, notably advertisements, on websites (e.g., [1]). Many websites providing content, such as news, outsource their advertising operations to large third-party ad networks, such as Google's DoubleClick. These networks embed tracking code into webpages across many sites providing the network with a more global view of each user's behaviors.

People are concerned about behavioral marketing on the web (e.g., [2]). To increase transparency and control, Google provides Ad Settings, which is "a Google tool that helps you control the ads you see on Google services and on websites that partner with Google" [3].

- Study of impact of different ad privacy settings
- Disclosing Gender resulted in fewer ads for high-paying jobs

PURDUE UNIVERSITY
Department of Computer Science

And it isn't just CS people who notice

"INTELLECTUAL FREEDOM AND RACIAL INEQUALITY AS ADDRESSED IN 'ALGORITHMS OF OPPRESSION'"

DR. SAFIYA NOBLE, Best-selling Author of *Algorithms of Oppression*
As Seen in *Wired*, *Time*, and Heard on NPR's *Science Friday*

Lecture 6–7 p.m.
Wednesday, Oct. 3, 2018
Fowler Hall | Stewart Center
30 minute Q&A following lecture
Free and open to the public

- **In an increasingly automated world, what IF AI tools punish the poor?**
- Feb. 13, 2019 Fowler Hall Purdue U.

31

PURDUE UNIVERSITY
Department of Computer Science

What are the reasons?

- Discrimination intentionally programmed into the system?
  - Let's hope not
- Historical bias in the training data?
  - May explain some, but not all
- Insensitivity on the part of developers?
  - Maybe
- Or perhaps we don't know (yet)?

PURDUE UNIVERSITY
Department of Computer Science

Conventional Wisdom: *It's the Training Data*

- "Data is frequently imperfect in ways that allow these algorithms to inherit the prejudices of prior decision makers."
  - Solon Barocas and Andrew Selbst, Big Data's Disparate Impact, *104 California Law Review 671* (2016)
- "Bias can easily creep into seemingly objective algorithms due to the selective nature of the training data"
  - Sidebar highlight in Jamie Griffith's The ineradicable bias at the heart of algorithm design, *The Panoply*, 2/15/19
- "We often shorthand our explanation of AI bias by blaming it on biased training data. The reality is more nuanced"
  - Karen Hao, This is how AI bias really happens—and why it's so hard to fix, *Technology Review* 2/14/19
  - Proceeds to discuss three ways that training data becomes biased (beyond historical bias)

Misconception

33

PURDUE UNIVERSITY
Department of Computer Science

Potential sources

- Historical bias in training data
  - Can we detect this?
- Feedback bias
  - Meth lab reports in Terre Haute
    - Increase police presence
  - Nearly 400 Meth labs in Terre Haute!
    - Is Terre Haute really the hotbed of Meth?
- "Tyranny of the majority"
  - Small populations deemed outliers
  - Algorithms effective "on average", but ignore rare cases
- Wrong objective function
  - Is accuracy the right measure?

PURDUE UNIVERSITY
Department of Computer Science

Credit Scoring using Decision Trees *(with Abhishek Sharma)*

- Experiment in Fairness using Statlog (German Credit Data) Data Set

  Data made available by Professor Dr. Hans Hofmann, Universität Hamburg via the UCI Machine Learning Repository

- Learn a decision tree from historical decisions
  - Data about credit applications
  - Decision made
    - *Better training data would be if loan was repaid…*
- Decision tree: model used to make future decisions
  - Goal is to make similar decisions to historical data

35

PURDUE UNIVERSITY
Department of Computer Science

So Where Is the Problem?

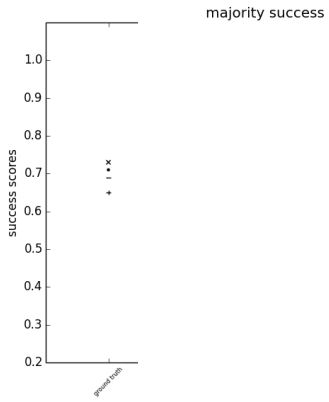
- We can show that some machine learning techniques should *reduce* bias from that in the training data
  - So why do we have so many examples of biased ML?
- **It isn't just the training data!**

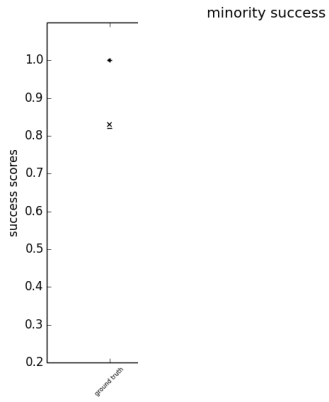
Myth: *Machines are Unbiased*

- Machine Learning can *introduce* bias against minority groups
  - Even when the minority is *advantaged*

49

Credit Dataset: Majority vs. Minority Positive Decisions

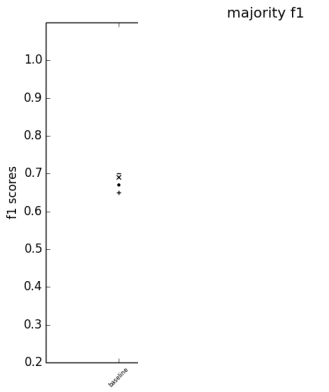


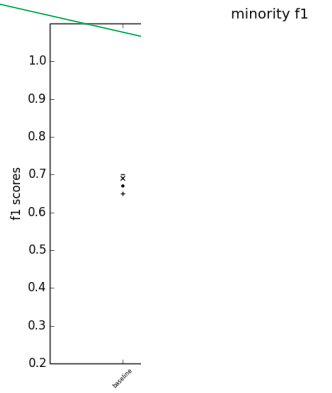


50

Credit Dataset: Majority vs. Minority Accuracy

Removing "bias"

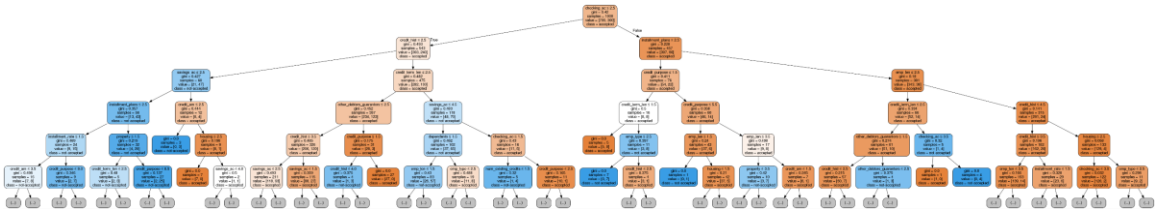


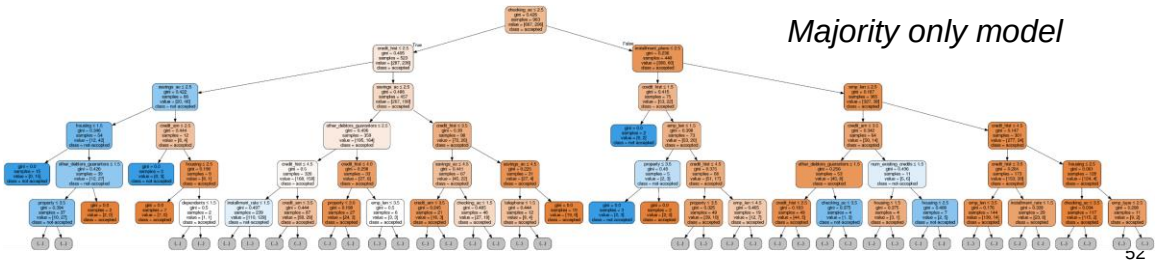


51

PURDUE UNIVERSITY
Department of Computer Science

Decision Tree



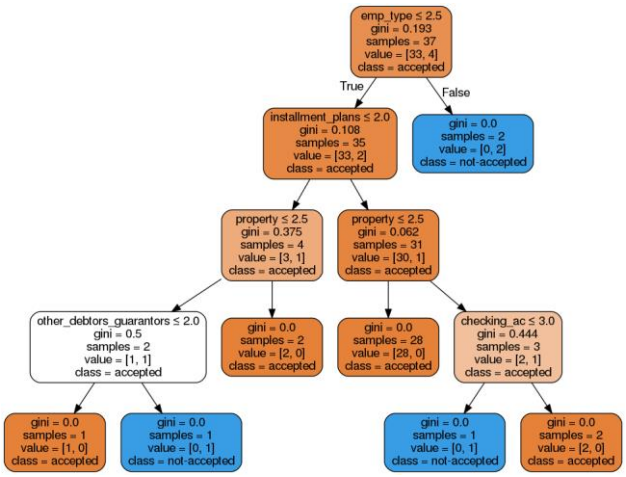


Majority only model

52

PURDUE UNIVERSITY
Department of Computer Science

Decision Tree: Minority Only Model



53

PURDUE UNIVERSITY
Department of Computer Science

Why is Machine Learning Introducing Bias?

- Key idea: ML typically optimizes for overall accuracy
- What is going on?
  - Distinct models that work best for majority, minority
  - Optimizing for global accuracy (revenue, …) selects model that works for majority
- Accurate / effective model for majority
  - But a bad model for the minority


Facebook Thinks Some Native American Names Are Inauthentic

Josh Sanburn @joshsanburn | Feb. 14, 2015

The social network is barring some Native Americans from logging in

If you're Native American, Facebook might think your name is fake.

The social network has a history of telling its users that the names they're attempting to use aren't real. Drag queens and overseas human rights activists, for example, have experienced error messages and problems logging in in the past.

Jörg Carstensen—AP

Some Native Americans say Facebook won't allow them to log in because their names are "inauthentic."

PURDUE UNIVERSITY
Department of Computer Science

Balance Training Data

- What if we get rid of majority/minority? (with Murat Kantarcioglu and Yan Zhou, UT Dallas)
- Augment training data with synthetic data
  - Generated to be similar to real data
- Synthetic data skewed to eliminate disparity in training data
  - Balance sizes of privileged/unprivileged groups
  - Balance positive/negative outcomes between groups

56

PURDUE UNIVERSITY
Department of Computer Science

What can we do?

- Detect discriminatory outcomes from machine learning
  - [Pedreschi08, Pedreschi09, Luong11, Ruggieri11]
- Relabel training samples
  - [Kamiran09, Zliobaite11, Kamiran11]
- Adjust scoring functions
  - [Calders10, Kamiran10]
- statistical parity
  - [Dwork12, Zemel13]

Myth: We Just Need Statistical Equality

PURDUE UNIVERSITY
Department of Computer Science

Multiple Measures:
Disparate Treatment vs. Disparate Impact

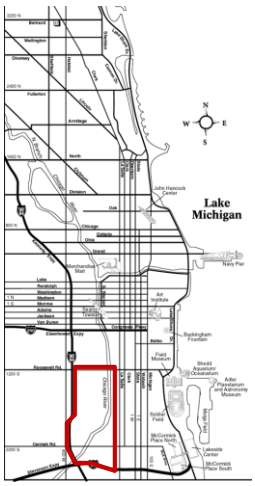
- Disparate treatment: Individuals from different groups treated differently
  - Otherwise identical individuals have different outcome based only on group membership
- Disparate impact: Outcomes different between different groups
  - No individuals are "the same"
  - Different outcomes for different groups, even if some other explanation
- Prior work largely relies on *using* special categories
  - This can qualify as disparate treatment

PURDUE UNIVERSITY
Department of Computer Science

Why Disparate Impact?

- Mortgage Redlining
  - Racial discrimination in home loans prohibited in US
  - Banks drew lines around high risk neighborhoods!!!
  - These were often minority neighborhoods
  - Result: Discrimination (redlining outlawed)

  What about data mining that "singles out" minorities?



PURDUE UNIVERSITY
Department of Computer Science

GDPR Requirement: Can't Use Certain Categories

- Article 22(4) Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

60

PURDUE UNIVERSITY
Department of Computer Science

Is Unbiased Training Data Enough?

- Rakin Haider: ML bias from unbiased data
- Assumptions:
  - Training data correct
  - Privileged and unprivileged groups of same size
  - Positive outcome probability same for both groups
- Difference
  - Different optimal models for the two groups
  - Optimal model for privileged group is higher accuracy

61

PURDUE UNIVERSITY
Department of Computer Science

Result: Biased Outcome

- Resource-scarce environment (e.g., selective college admissions): Optimal accuracy global model favors privileged class
  - This wasn't true in the training data
- Analysis based on Bayesian model
  - Presumably "good" practical ML will do the same
  - Demonstrated on a variety of real-world classifiers
    - *Including some explicitly designed to reduce bias*
- Reflects a type of **Systemic Bias**

62

PURDUE UNIVERSITY
Department of Computer Science

Ethics Issues for Data Mining & ML
What's the Problem?

- Privacy
  - Training data
  - Allowed uses
- Fairness
  - Inequitable outcomes
  - Variance in accuracy
- Data inaccuracy
- Explainability
- Redress
  - What if someone disputes results?

68

PURDUE UNIVERSITY
Department of Computer Science

Transparency

- Analyze and explain AI decision process
  - Very difficult
  - Likely only understandable to technology and domain experts
- Analyze and explain a decision
  - Input data analysis
  - Static explanation
  - Design/Code review and statistical analysis
  - Sensitivity analysis
  - Reverse-engineering the model

70

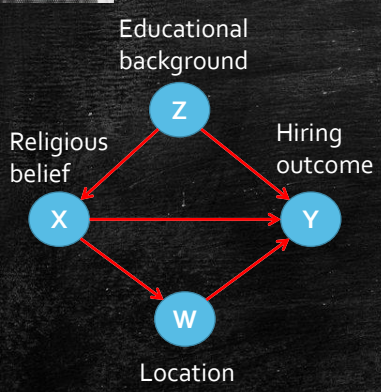
PURDUE UNIVERSITY
Department of Computer Science

GDPR Requirement: Transparency

- Article 13(2)(f), 4(2)(g): the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.
- Article 22(1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
- Article 22(4) Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

71

Static Explanation through Causal Reasoning
(Junzhe Zhang and Elias Bareinboim AAAI'18)



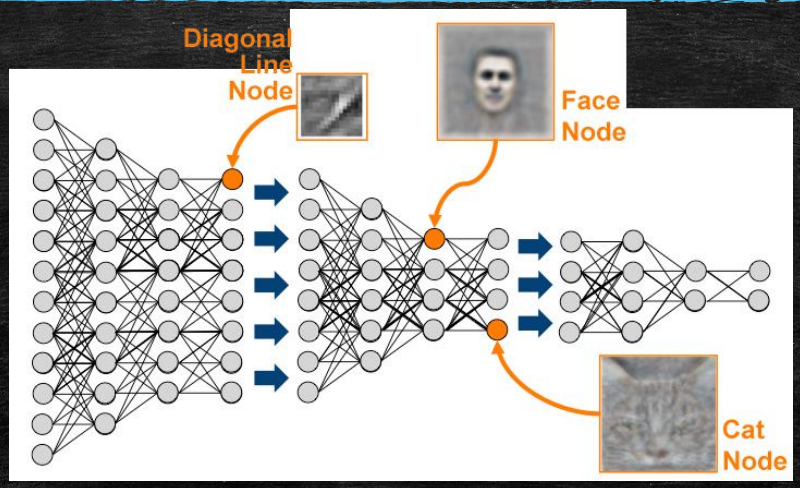
- The data analysis reveals that the total variation
$$E[Y|X=1] - E[Y|X=0] \ll 0$$
i.e., applicants of faith has lower chance of being hired.
- A frustrated applicant sues the company, claiming the disparity is due to:
  - Direct discrimination: the direct path $X \rightarrow Y$.
  - Indirect discrimination: the indirect path $X \rightarrow W \rightarrow Y$.
- The company argues the disparity is due to:
  - Difference in educational background: the spurious path $X \leftarrow Z \rightarrow Y$.

- Challenge: We do not have access to the code of the decision-making system (or the brains of the HR personnel in charge of hiring), so how to determine who is telling the truth?

Fairness in Decision-Making, Zhang and Bareinboim, AAAI'18.

72

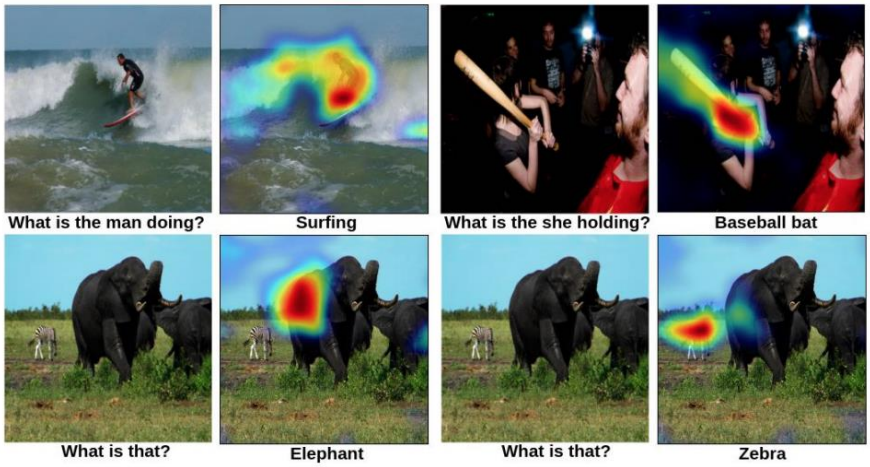
Reverse Engineering the Model
Back to Neural Nets



75

The University of Texas at Austin — WHAT STARTS HERE CHANGES THE WORLD

Visual Explanation

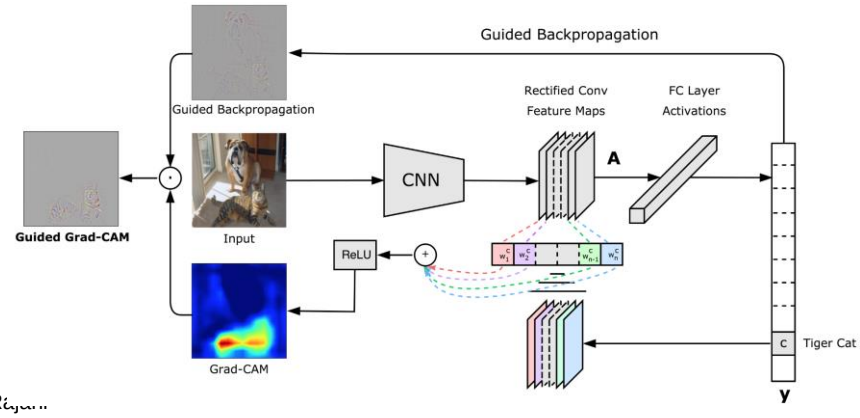


Dr. Nazneen Rajani

76

Generating Visual Explanation

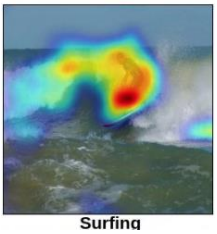
- *GradCAM* (Selvaraju et al., 2017) is used to generate heat-map explanations.

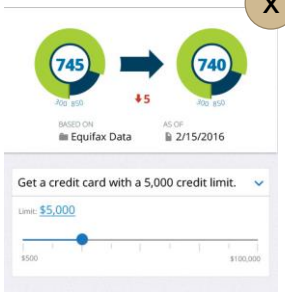


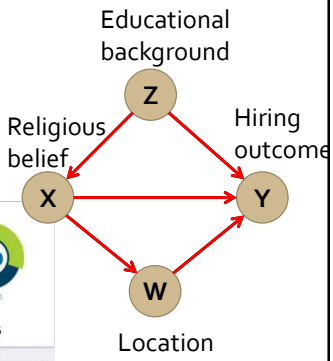
PURDUE UNIVERSITY
Department of Computer Science

Are Explanations Accurate?

- Do these explanations really capture how decisions are made?
  - Sensitivity Analysis, Causal Reasoning
    - Explain outcome, not process
  - Heat maps
    - maybe?
- But does it matter?







PURDUE UNIVERSITY
Department of Computer Science

Emotional vs. Rational Decision-Making

- Humans have been shown to be emotional in their decision making
  - fMRI analysis of how decisions are made *(De Martino, Kumaran, Seymour, Dolan, Science 2006)*
- We rationalize our decisions
  - Explanations justify why we the decisions are good, not how we make them
- Is this good enough for explaining AI?
  - *Does this qualify as making ethical decisions?*

79

PURDUE UNIVERSITY
Department of Computer Science

Ethics Issues for Data Mining & ML *What's the Problem?*

- Privacy
  - Training data
  - Allowed uses
- Fairness
  - Inequitable outcomes
  - Variance in accuracy
- Data inaccuracy
- Explainability
- Redress
  - What if someone disputes results?

81

PURDUE UNIVERSITY
Department of Computer Science

Top Ethical Issues
As presented at 2016 WEF

1. Unemployment
2. Distribution of machine-created wealth
3. Impact on human behavior/interaction
4. Guarding against mistakes
5. AI bias
6. Safety from adversaries
7. Protect against unintended consequences
8. How do we stay in control?
9. Robot rights

87

PURDUE UNIVERSITY
Department of Computer Science

Ethical Issues:
AI Safety

- Multiple issues
  - Mistakes
  - Unintended consequences
  - Protection from adversaries
- Can we guarantee certain outcomes?
  - Rule out bad outcomes?

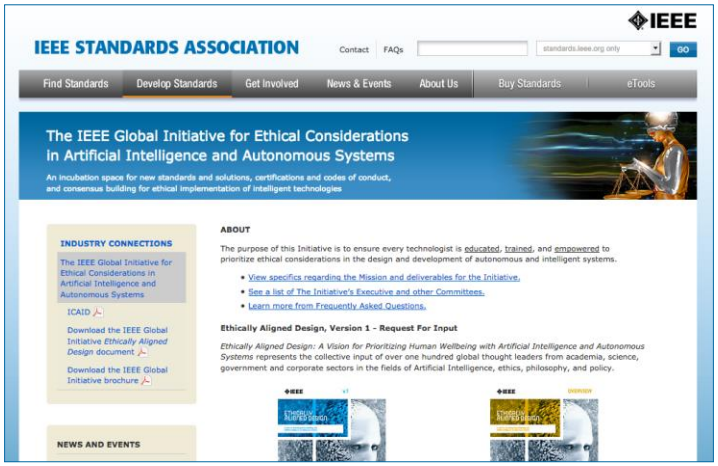
88

PURDUE UNIVERSITY
Department of Computer Science

Can We Trust Machine Learning?

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song. Robust Physical-World Attacks on Deep Learning Visual Classification. *Computer Vision and Pattern Recognition (CVPR 2018)*

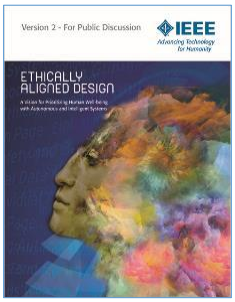
What do we do about it?
Standards and Best Practices



IEEE STANDARDS ASSOCIATION IEEE

Ethically Aligned Design

A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems



Version 2

- Launched December 2017 as a Request for Input
- Created by over 250 Global A/IS & Ethics professionals, in a bottom up, transparent, open and increasingly globally inclusive process
- Incorporates over 200 pages of feedback from public RFI and new Working Groups from China, Japan, Korea and more
- Thirteen Committees / Sections
- Contains **over one hundred twenty** key Issues and Candidate Recommendations

https://ethicsinaction.ieee.org/

IEEE STANDARDS ASSOCIATION IEEE

IEEE P70xx Standards Projects

IEEE P7000: Model Process for Addressing Ethical Concerns During System Design
IEEE P7001: Transparency of Autonomous Systems
IEEE P7002: Data Privacy Process
IEEE P7003: Algorithmic Bias Considerations
IEEE P7004: Child and Student Data Governance
IEEE P7005: Employer Data Governance
IEEE P7006: Personal Data AI Agent Working Group
IEEE P7007: Ontological Standard for Ethically Driven Robotics and Automation
IEEE P7008: Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
IEEE P7009: Fail-Safe Design of Autonomous and Semi-Autonomous Systems
IEEE P7010: Wellbeing Metrics Standard for Ethical AI and Autonomous Systems
IEEE P7011: Process of Identifying and Rating the Trustworthiness of News Sources
IEEE P7012: Standard for Machines Readable Personal Privacy Terms

IEEE STANDARDS ASSOCIATION IEEE

95

Related AI standards activities

- British Standards Institute (BSI) – BS 8611 *Ethics design and application of robots*

- **ISO/IEC JTC 1/SC 42 Artificial Intelligence**
  - ***SG 1 Computational approaches and characteristics of AI systems***
  - ***SG 2 Trustworthiness***
  - ***SG 3 Use cases and applications***
  - ***WG 1 Foundational standards***

- Jan 2018 China published "Artificial Intelligence Standardization White Paper."

IEEE STANDARDS ASSOCIATION IEEE

PURDUE UNIVERSITY
Department of Computer Science

General Guidelines: FIPPs

Fair Information Practice Principles

- Transparency
  - Organizations should be transparent and notify individuals
- Individual Participation
  - Organizations should involve the individual in the process of using PII
- Purpose Specification
  - Organizations should specifically articulate the authority that permits the collection of PII
- Data Minimization
  - Organizations should only collect PII that is directly relevant and necessary
- Use Limitation
  - Organizations should use PII solely for the purpose(s) specified in the notice
- Data Quality and Integrity
  - Organizations should, to the extent practicable, ensure that PII is accurate, relevant, timely, and complete.
- Security
  - Organizations should protect PII (in all media) through appropriate security safeguards
- Accountability and Auditing
  - Organizations should be accountable for complying with these principles

97

NATIONAL STRATEGY FOR TRUSTED IDENTITIES IN CYBERSPACE - Appendix A

Quiz: Which of the FIPPs were violated in the Criminal Recidivism case?

- Propublica reporters analyzed the COMPAS recidivism software on data from Broward County, Florida and found that it discriminated against people of color
- *Note: Equivant (developer of COMPAS) did an analysis on the same data and concluded it wasn't discriminatory – they used a different definition*

A. Transparency
B. Individual Participation
C. Purpose Specification
D. Data Minimization
E. Use Limitation
F. Data Quality and Integrity
G. Security
H. Accountability and Auditing

98

Quiz: Which of the FIPPs were violated by Cambridge Analytica?

- Cambridge Analytica used a Facebook app to capture the profile information of users of the app and their friends
- This was used for political analysis to target individualized messages to voters in the 2016 US Presidential Election

  This was considered bad enough that Mark Zuckerberg was called to testify before Congress!

A. Transparency
B. Individual Participation
C. Purpose Specification
D. Data Minimization
E. Use Limitation
F. Data Quality and Integrity
G. Security
H. Accountability and Auditing

99

PURDUE UNIVERSITY
Department of Computer Science

Ethical Reasoning

- Ethical: Of or relating to moral principles
- Moral (of an action): having the property of being right or wrong, voluntary or deliberate and therefore open to ethical appraisal
- Ethical Reasoning in the context of AI (NSW Government):
  - A process of identifying ethical issues and weighing multiple perspectives to make informed decisions
  - Not about knowing right from wrong, but being able to think about and respond to a problem fairly, justly, and responsibly

104

PURDUE UNIVERSITY
Department of Computer Science

Some suggestions

- Attend relevant talks
  - CS colloquium series (lists.purdue.edu – cs-colloq)
  - www.purdue.edu/critical-data-studies
- Data Ethics courses (a few)
  - ILS 23000: Data Science and Society: Ethical, Legal, Social Issues
  - PHIL 20700: Ethics for Technology, Engineering, and Design
  - PHIL 20800: Ethics of Data Science

105