PURDUE UNIVERSITY | Department of Computer Science

CS47300: Web Information Search and Management

Search Ethics: Bias

Prof. Chris Clifton

21 October 2020

PURDUE UNIVERSITY
Department of Computer Science

Ethics Issues for Web Search
What's the Problem?

- Privacy
  - Query
  - Pages clicked
  - Profiles
- Inappropriate search results
  - Children
  - "Picking" what you want people to see
  - Racial/Gender/Ethnic/… bias

38

PURDUE UNIVERSITY
Department of Computer Science

US Law: COPPA
Children's Online Privacy Protection Rule

- COPPA restricts:
  - Enabling a child to make personal information publicly available in identifiable form
  - Passive tracking of a child online
  - Collecting children's information for profiling and behavioral advertising
  - Requiring personal information to participate in online games/activities
- Child Online Protection Act
  - Would have restricted internet transmission of material harmful to minors
  - Struck down as unconstitutional

39

PURDUE UNIVERSITY
Department of Computer Science

Restriction on Search Results

- German law prohibits hate speech (*Volksverhetzung*)
  - Includes glorifying National Socialism, Holocaust Denial
  - Has been used to require Google to remove sites from search results
- United Kingdom (and others) restricts certain searches
  - Blacklisted searches must return no results

40

PURDUE UNIVERSITY
Department of Computer Science

Filter Bubbles

- Search engine goal
  - Satisfy your information need?
  - Sell advertising?
  - *Keep you coming back!*
- Give you what you want to see

- What do you want to see?
  - Things that match your query
  - What other people like
    - Pagerank
  - What you've liked in the past
    - Profiling (we'll discuss this later)
  - What others like you like
    - Collaborative filtering

41

PURDUE UNIVERSITY
Department of Computer Science

Filter Bubbles: Problem

- Your goal (hopefully): Satisfy information need
  - Technologies customize this to what you are predicted to like
  - See only subset of information
    - Typically the same subset
- Outcome: Myopic view of the world
  - Not the best information, but information that matches predictions
  - Or, matches your expectations
- Personal and societal implications

42

PURDUE UNIVERSITY
Department of Computer Science

Experiment: Disable Profiling

- Turn off profiling for a week
  - See if you notice a difference
- It isn't easy to do
  - Privacy settings
    - Hard to find, limited capabilities
  - Cookies
    - Turn them off entirely and a lot of sites break
  - Web beacons
    - Install blockers – but be careful, not all are reputable

PURDUE UNIVERSITY | Department of Computer Science

Big Data Ethics: Detecting Bias in Data Collection, Algorithmic Discrimination and 'Informed Refusal'

Chris Clifton, Daniel Kelly, Kendall Roark

PURDUE UNIVERSITY
Department of Computer Science

Discrimination in AI: What's all the fuss?

Facebook Thinks So...

Amazon scraps secret A... that sho...

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

...ome 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?**
- Prof. Virginia Eubanks, U. Albany, Feb. 13, 2019 Fowler Hall Purdue U.

56

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

58

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

60

PURDUE UNIVERSITY
Department of Computer Science

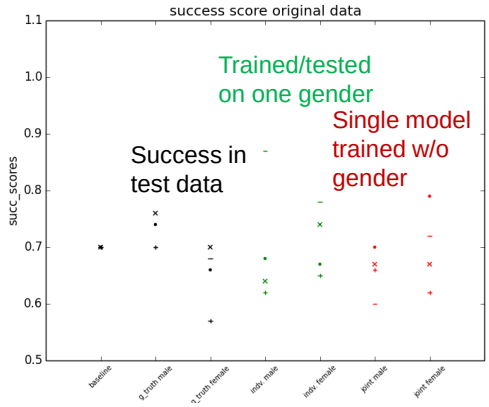
Evaluating Impact of Biased Data

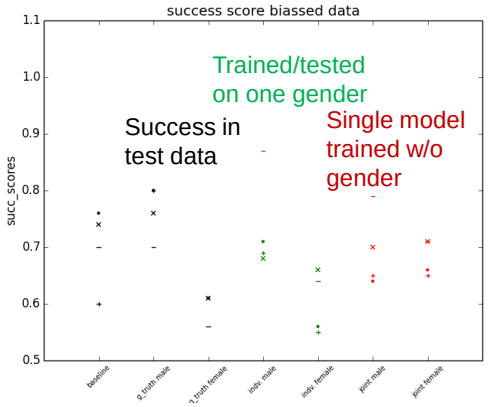
- Prior work has discovered little gender bias in this dataset
  - Pedreschi et al., Mancuhan & Clifton '14
  - Some disparity, but well-explained by other factors
- What happens if we *induce* gender bias?
  - Does the learned model show bias?
- Trained models on original data, data with x% of decisions changed to favor males over females
- Baseline: "all data" (including Gender)
  - Gender-specific models
  - "Gender-blind" model

61

PURDUE UNIVERSITY
Department of Computer Science

Gender Bias: Success Rate

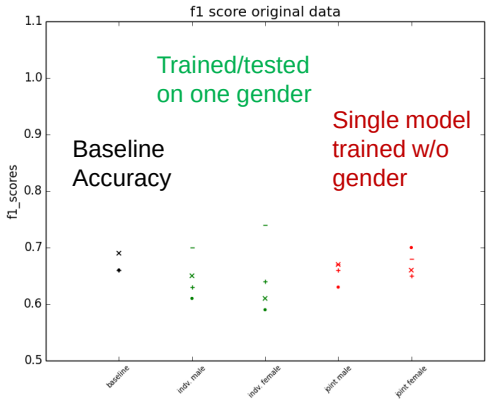


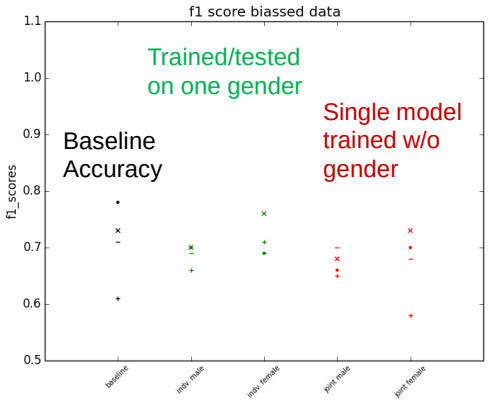


62

PURDUE UNIVERSITY
Department of Computer Science

Gender Bias: Accuracy





63

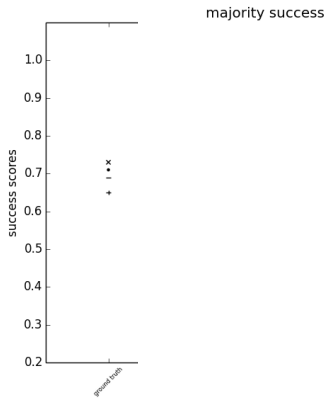
PURDUE UNIVERSITY
Department of Computer Science

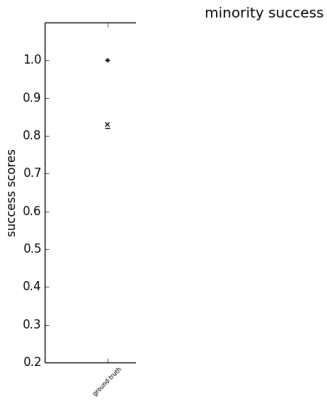
Potential sources

- 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 Dataset: Majority vs. Minority Positive Decisions





78

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

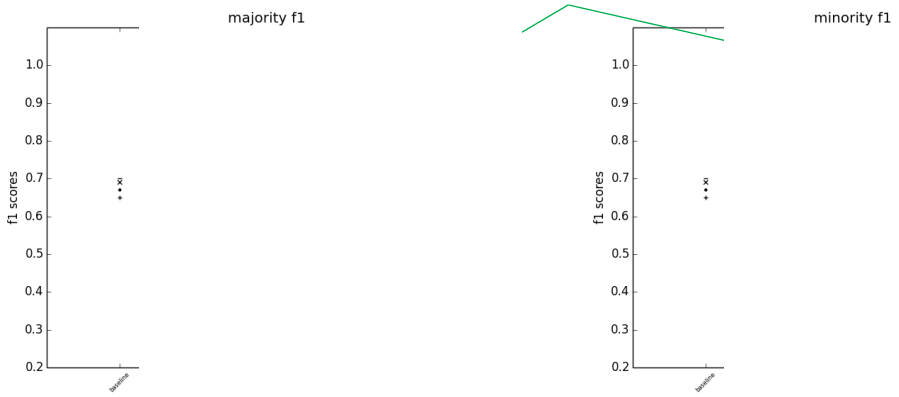
GDPR Requirement: Transparency

- 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.

PURDUE UNIVERSITY
Department of Computer Science

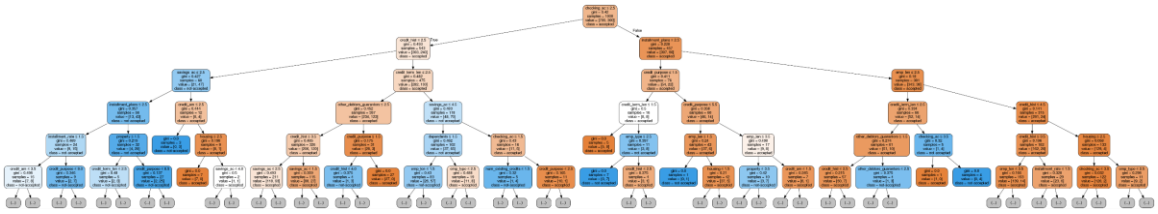
Credit Dataset: Majority vs. Minority Accuracy

Removing "bias"

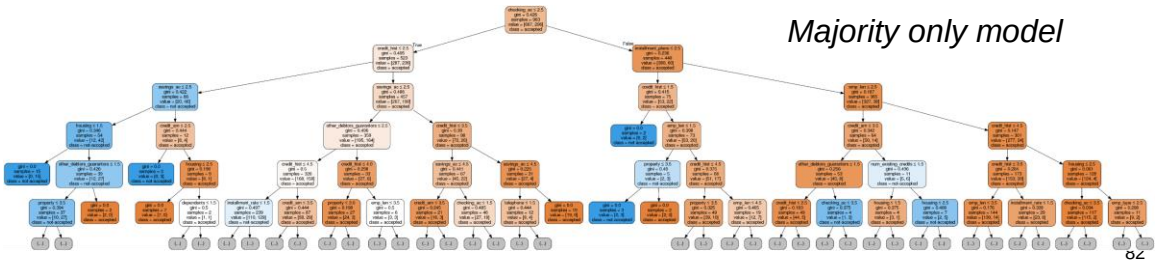


PURDUE UNIVERSITY
Department of Computer Science

Decision Tree



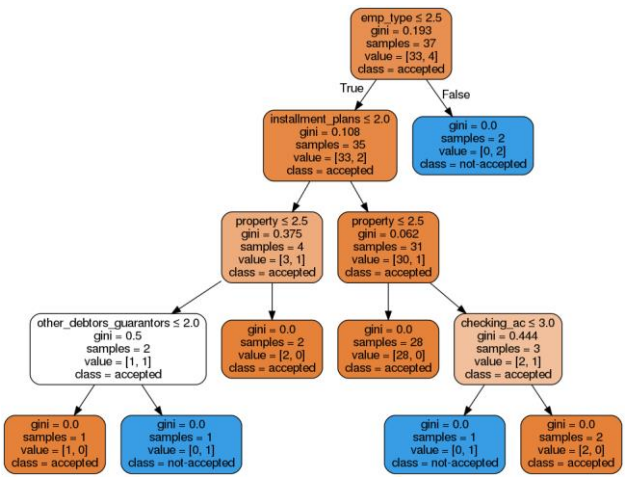
Majority only model



82

PURDUE UNIVERSITY
Department of Computer Science

Decision Tree: Minority Only Model



83

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.

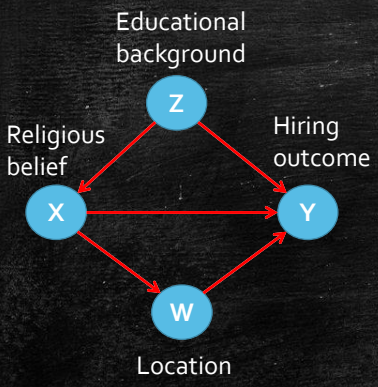
85

Fairness in Decision-Making -- The Causal Explanation Formula *(Junzhe Zhang and Elias Bareinboim AAAI'18)*

- Goal: Determine the specific mechanisms by which the protected attribute brings about change in the outcome variable (decision), without having a priori knowledge about the decision-making mechanisms.
- Results: First, we introduced a new family of measures, based on causal inference, capable of detecting these mechanisms uniquely. We further derived the causal explanation formula, which allows one, for the first time, to decompose the observed discrimination in the specific discriminatory pathways present in the underlying decision-making process.
- Vision: Develop a principled framework to understanding and explaining fairness problems in automated decision-making systems, which involves the challenge of translating unobserved human biases embedded in past decisions (present in the training data) into transparent causal quantities.

100

Example: Discrimination in Hiring



- 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 $X \leftarrow Z \rightarrow Y$.

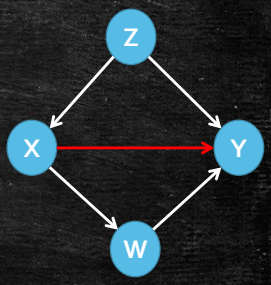
- Challenge: We do not have access to the code of the decision-making system (or the brains of the HR personnel in charge of hiring), so how to determine who is telling the truth?

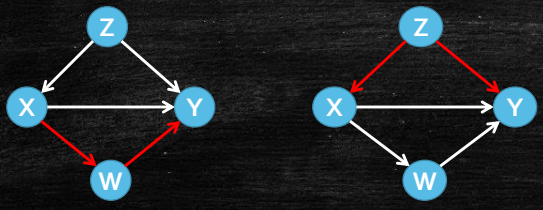
Fairness in Decision-Making, Zhang and Bareinboim, AAAI'18.

101

Novel Counterfactual Measures for Path-Specific Effects

- The direct effect of the protected attribute X on the outcome Y is given by the counterfactual quantity:
$$DE_{x_0,x_1}(Y|x) = E\left[Y_{x_1,W_{x_0}}|x\right] - E\left[Y_{x_0}|x\right]$$
- (See paper for formal semantics and interpretation.)
- The counterfactual indirect (left) and spurious (right) effects can be formalized in similar fashion. Graphically:



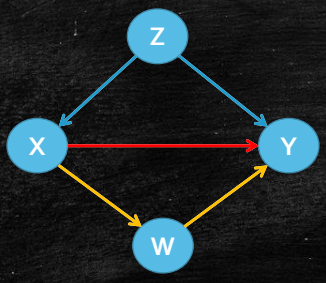


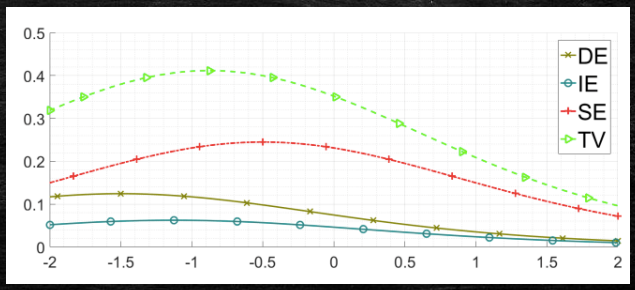
Fairness in Decision-Making, Zhang and Bareinboim, AAAI'18.

102

Quantifying Discrimination – The Causal Explanation Formula

Theorem. The total variation (TV), direct, indirect , and spurious effects satisfy the causal explanation formula, i.e.: $TV_{x_0,x_1}(Y) = \underbrace{DE_{x_0,x_1}(Y|x_0)}_{Red} - \underbrace{IE_{x_1,x_0}(Y|x_0)}_{Yellow} - \underbrace{SE_{x_1,x_0}(Y)}_{Blue}$





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

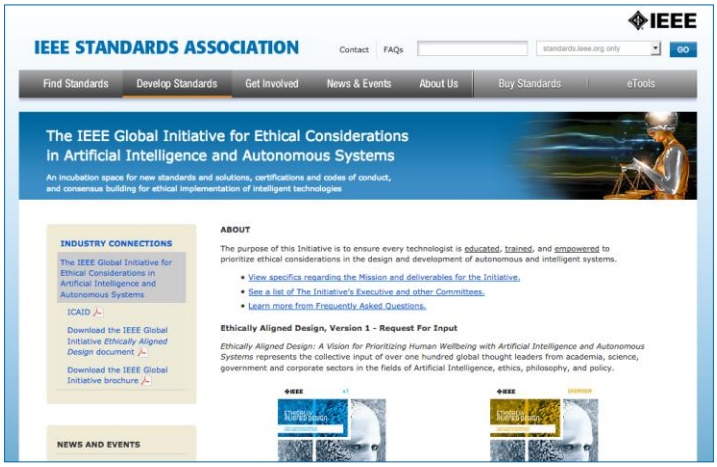
103

PURDUE UNIVERSITY
Department of Computer Science

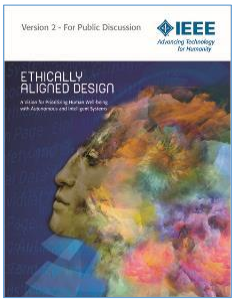
Ideas for the Future

- Tests for Bias?
  - Or perhaps just *potential* bias?
  - ethicstoolkit.ai
- Fundamental changes in machine learning?
  - Objective functions other than accuracy
- IEEE-SA P7003: Standard for Algorithmic Bias Considerations
  - *Work in Progress*
- Understand distinction between Bias and Personalization (supported by the Mellon Foundation):
  - What determines if a recommendation is "Biased" or "Personalized"
  - Explored Participatory Design to elicit issues
  - *Joint work with Kendall Roark (Data Ethicist, Purdue Libraries) and Daniel Kelly (Purdue Philosophy Dept.)*

What do we do about it?
Standards and Best Practices



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 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

http://sites.ieee.org/sagroups-7003/

P7003 foundational sections

- Taxonomy of Algorithmic Bias
- Legal frameworks related to Bias
- Psychology of Bias
- Cultural aspects

P7003 algorithm development sections

- Algorithmic system design stages
- Person categorization and identifying affected population groups
- Assurance of representativeness of testing/training/validation data
- Evaluation of system outcomes
- Evaluation of algorithmic processing
- Assessment of resilience against external manipulation to Bias
- **Documentation of criteria, scope and justifications of choices**

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."