Protected Attributes and ‘Fairness through Unawareness’

Exploring Fairness in Machine Learning

Mike Teodorescu
Assistant Professor of Information Systems, Boston College
Visiting Scholar, MIT

MITD-Lab USAID FROM THE AMERICAN PEOPLE CITE

Attributes Associated with Social Bias

Certain individual attributes are tied to social bias (often referred to as ‘protected attributes’) :

- race;
- religion;
- national origin;
- gender;
- marital status;
- age;
- socioeconomic status.

Example of Laws In the US

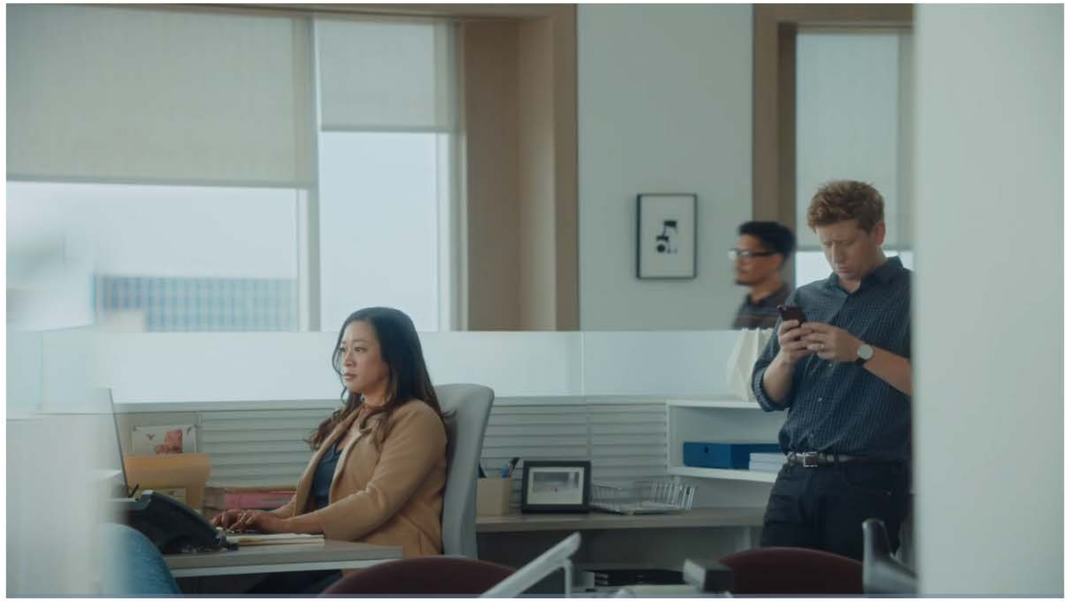
In the US, there have been laws that prohibit discriminating on the basis of these attributes in applications like housing, credit lending, and employment:

- Penalties for discrimination in housing (US Fair Housing Act)
- Hiring (the collection of laws also known as Federal Equal Employment Opportunity – Civil Rights Act Title VII 1964, EPA 1963, ADEA 1967, ADA 1990, Rehabilitation Act 1973, Civil Rights Act 1991, GINA 2008).
- Lending (Equal Credit Opportunity Act)

Regardless of legal framework, machine learning has the potential to unintentionally embed bias.

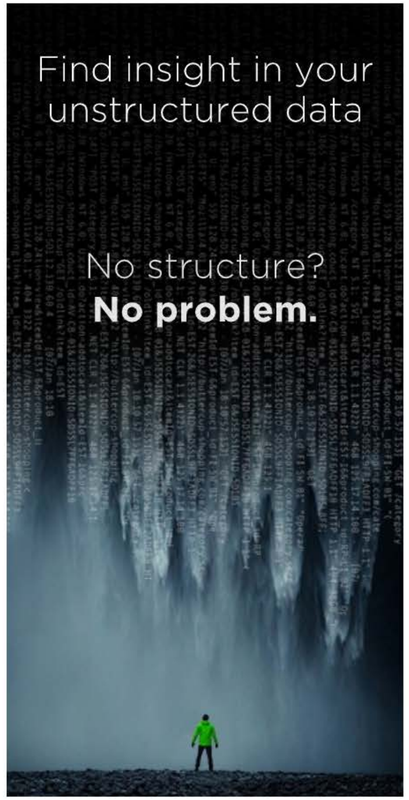
MPW • AMAZON

Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women



By DAVID MEYER October 10, 2018

Machine learning, one of the core techniques in the field of artificial intelligence, involves teaching automated systems to devise new ways of doing things, by feeding them reams of data about the subject at hand. One of the big fears here is that biases in that data will simply be reinforced in the AI systems —and Amazon seems to have just provided an excellent example of that phenomenon.



You May Like by Outbrain

Born After 1943? You Could

© Fortune Media. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

source:
https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/

IAN WALDIE/GETTY IMAGES

Tech Policy / AI Ethics

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by Karen Hao Jan 21, 2019

© MIT Technology Review. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

source:
https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/

The New York Times

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

170

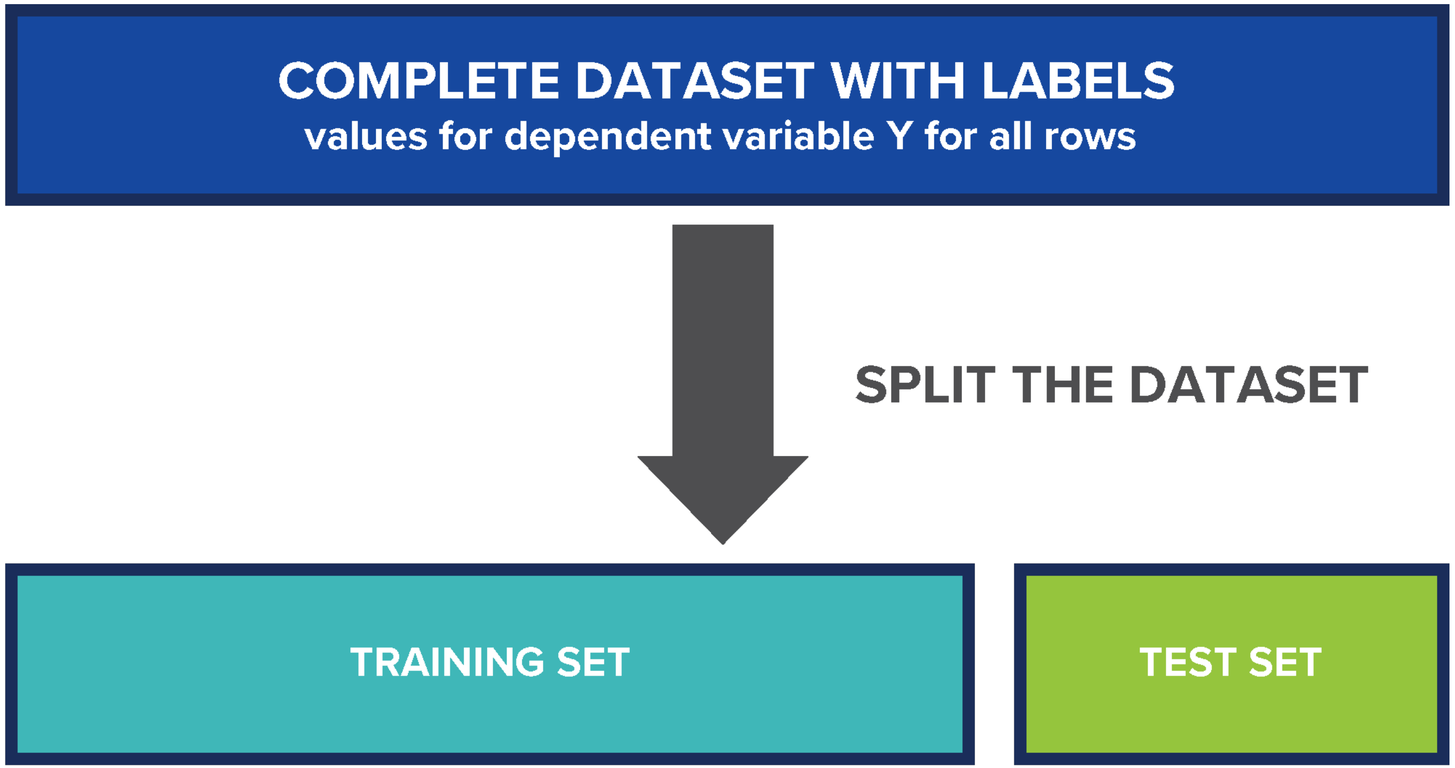
WASHINGTON — The Department of Housing and Urban Development sued Facebook on Thursday for engaging in housing discrimination by allowing advertisers to restrict who is able to see ads on the platform based on characteristics like race, religion and national origin.

In addition to targeting Facebook's advertising practices, the housing department, known as HUD, claims in its lawsuit that the company uses its data-mining practices to determine which of its users are able to view housing-related ads. On both counts, the agency said, Facebook is in violation of the federal Fair Housing Act.

© NY Times. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

source:
https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html

Train, Test



Key Point: You should randomly sample the training set and the test set

Fairness Starts with the Training Set

- The training set can carry the biases of the people labeling the data
- Bad training data => bad prediction
- The training data may not be representative of all the groups
- Hidden correlations in input data
- Individuals may misremember past situations - selective perception (Dearborn & Simon, 1958)

Base Case: Fairness Through Unawareness

- The default fairness method in machine learning is fairness-through-unawareness
- Fairness-through-unawareness refers to leaving out of the model protected social attributes such as gender, race, and other characteristics deemed sensitive
- However, ignoring meaningful group differences does not erase inequality but instead can perpetuate it.

Failures of Fairness through Unawareness

- When race, gender, and other sensitive variables are treated as protected, other variables such as college attended, hometown, or various resume indicators that remain unprotected may still be highly correlated with the protected attributes.

- For example, researchers at Carnegie Mellon University revealed that gender, a protected attribute, caused an unintentional change in Google's advertising system such that ad listings targeted for users seeking high-income jobs were presented to men at nearly six times the rate they were presented to women (Datta et al., 2015).

Review Questions

- What are the sensitive attributes in the context in which you work?
- Do you think the current list of protected attributes is exhaustive?
- What is “fairness through unawareness”?
- What variables might lead to biased predictions for a machine learning hiring system in your country?
- What are some risks to an organization choosing “unawareness”?

Acknowledgments

- Joint study with Professors Lily Morse and Gerald Kane (Boston College) and Yazeed Awwad (Research Assistant, MIT D-Lab).

- The authors thank USAID-MIT Grant AID-OAA-A-12-00095 “Appropriate Use of Machine Learning in Developing Country Contexts” and the Carroll School of Management at Boston College for research funding. I thank Research Assistant Mariana Paredes and Daniel Brown (HBS), Dr. Daniel Frey (MIT), Dr. Aubra Anthony (USAID), Dr. Shachee Doshi (USAID), Dr. Amy Paul (USAID), Dr. Craig Jolley (USAID), Dr. Rich Fletcher (MIT), Amit Gandhi (MIT), Lauren McKown (MIT), Kendra Leith (MIT), Nancy Adams (MIT), and Dr. Sam Ransbotham (BC) for help with this work.

References

- Abdi, H. (2007). The Kendall rank correlation coefficient. Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA, pp.508-510.
- Agrawal, A., Gans, J. and Goldfarb, A. (2018). Prediction machines: the simple economics of artificial intelligence. Harvard Business Press, 195-206.
- Ajunwa, Ifeoma, The Paradox of Automation as Anti-Bias Intervention (forthcoming). Cardozo Law Review.
- Angst, C., Agarwal, R. (2009). Adoption of Electronic Health Records in the Presence of Privacy Concerns: The Elaboration of Likelihood Model and Individual Persuasion. MIS Quarterly, 33, 339-370.
- Angst, C. (2009). Protect My Privacy or Support the Common-Good? Ethical questions about electronic health information exchanges. Journal of Business Ethics, 90, 169-178.
- Apfelbaum, E. P., Pauker, K., Sommers, S. R., & Ambady, N. (2010). In Blind Pursuit of Racial Equality? Psychological Science, 21(11), 1587-1592.
- Ashcraft, C., McLain, B. and Eger, E. (2016). Women in tech: The facts. National Center for Women & Technology (NCWIT).
- Datta, A., Tschantz, M. C., &Datta, A. (2015). Automated Experiments on Ad Privacy Settings. Proceedings on Privacy Enhancing Technologies, 1, 92-112.
- Datta, A.,Fredrikson, M., Ko, G.,Mardziel, P., Sen, S. 2017. Proxy non-discrimination in data-driven systems. arXiv preprint arXiv:1707.08120.
- Dearborn, D.C., & Simon, H.A. (1958). Selective perception: A note on the departmental identifications of executives. Sociometry, 21, 140-144.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. ITCS '12 Proceedings of the 3rd Innovations in Theoretical Computer ScienceConference, 214-226. Cambridge, Massachusetts — Jan 08 - 10, 2012

Thank you

Mike Teodorescu
Assistant Professor of Information Systems, Boston College
Visiting Scholar, MIT

hmteodor@mit.edu

MITD-Lab USAID FROM THE AMERICAN PEOPLE CITE

MIT OpenCourseWare
https://ocw.mit.edu

RES.EC-001 Exploring Fairness in Machine Learning

Spring 2019

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.