AI

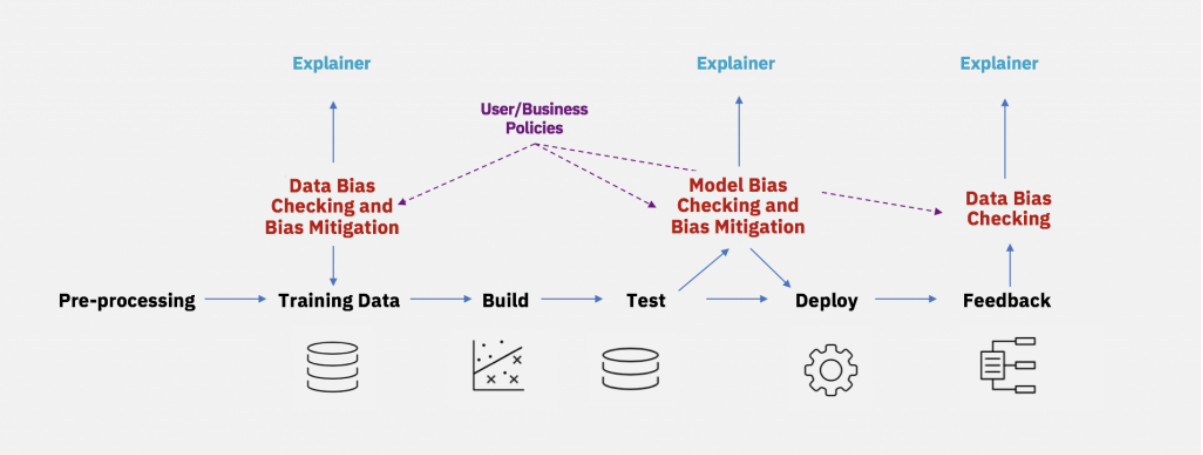
Introducing AI Fairness 360

- September 19, 2018 | Written by: Kush R. Varshney

Categorized: AI

Share this post:

We are pleased to announce AI Fairness 360 (AIF360), a comprehensive open-source toolkit of datasets and machine learning models, and state-of-the-art algorithms to mitigate such bias. W it to help engender trust in AI and make the world more equitable for all.



Mitigating bias throughout the AI lifecycle

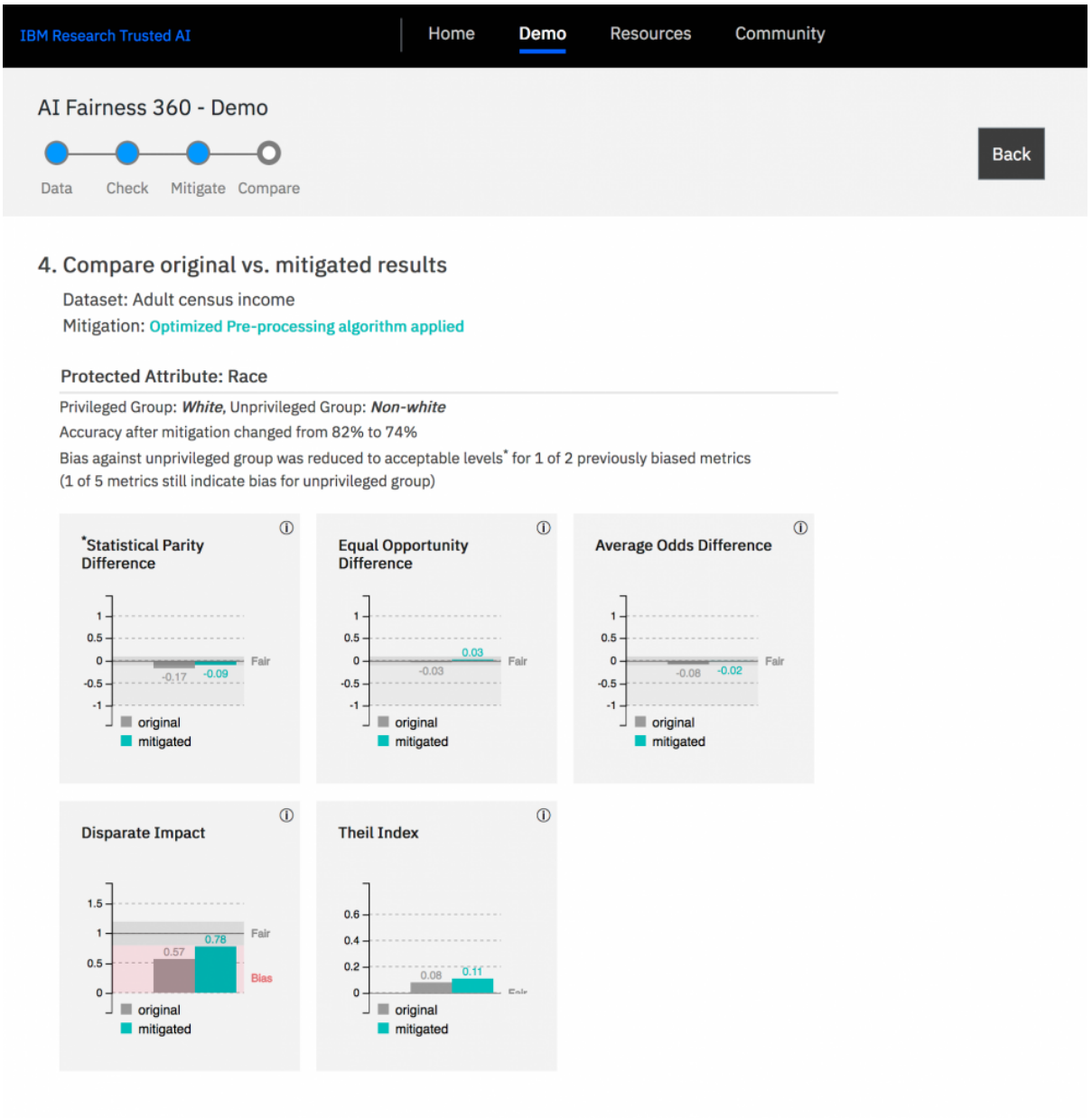
Machine learning models are increasingly used to inform high-stakes decisions about people. A nature, is always a form of statistical discrimination, the discrimination becomes objectionable groups at systematic advantage and certain unprivileged groups at systematic disadvantage. Bi prejudice in labels or under-/over-sampling, yields models with unwanted bias.

This initial release of the AIF360 Python package contains nine different algorithms, developed

research community, to mitigate that unwanted bias. They can all be called in a standard way, v
paradigm. In this way, we hope that the package is not only a way to bring all of us researchers

focus on industrial usability, and its software engineering.

AIF360 contains three tutorials (with more to come soon) on credit scoring, predicting medical
images by gender. I would like to highlight the medical expenditure example; we've worked in t
health insurance clients (without explicit fairness considerations), but it has not been considere
before. (For background, here are two papers describing our earlier applied data science work i



AI Fairness 360 interactive experience

AIF360 is not just a Python package. It is also an interactive experience that provides a gentle i
capabilities of the toolkit. Being a comprehensive set of capabilities, it may be confusing to figu
are most appropriate for a given use case. To help, we have created some guidance material tha

Our team includes members from the IBM India Research Lab and the T. J. Watson Research Ce
the toolkit as a summer project this year. We are a diverse lot in terms of national origin, scienti

One of the reasons we decided to make AIF360 an open source project as a companion to the a
encourage the contribution of researchers from around the world to add their metrics and algor
AIF360 becomes the hub of a flourishing community.

The currently implemented set of metrics and algorithms are described in the following list of p

Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, an
Processing for Discrimination Prevention," Conference on Neural Information Processing Sy

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkata
Removing Disparate Impact," ACM SIGKDD International Conference on Knowledge Discov

Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of Opportunity in Supervised Learning
Processing Systems, 2016.

Faisal Kamiran and Toon Calders, "Data Preprocessing Techniques for Classification withou
Information Systems, 2012.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang, "Decision Theory for Discrimination-Awa
Conference on Data Mining, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma, "Fairness-Aware Class
Regularizer," Joint European Conference on Machine Learning and Knowledge Discovery in

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger, "On Fairn
Neural Information Processing Systems, 2017.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian We
Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfa
SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018.

Richard Zemel, Yu (Ledell) Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork, "Learnir
Conference on Machine Learning, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell, "Mitigating Unwanted Biases with A
Conference on Artificial Intelligence, Ethics, and Society, 2018.

---

[1]Some of the excellent repositories are Aequitas, Audit-AI, FairML, Fairness Comparison, Fairne
Themis-ML.

[2]AIF360 team members are Rachel Bellamy, Kuntal Dey, Mike Hind, Sam Hoffman, Stephanie H
Jacquelyn Martino, Sameep Mehta, Saška Mojsilović, Seema Nagar, Karthi Natesan Ramamurth

Sattigeri, Moninder Singh, Kush Varshney, Dakuo Wang, and Yunfeng Zhang.

Principal Research Staff Member and Manager, IBM Research

AI | Artificial Intelligence | trusted AI

Previous Post

Helping to Improve Medical Image Analysis with Deep Learning

Trust an

More AI stories