**AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias**

# **Citation:** [Rachel K. E. Bellamy](https://arxiv.org/search/cs?searchtype=author&query=Bellamy%2C+R+K+E), [Kuntal Dey](https://arxiv.org/search/cs?searchtype=author&query=Dey%2C+K), [Michael Hind](https://arxiv.org/search/cs?searchtype=author&query=Hind%2C+M), [Samuel C. Hoffman](https://arxiv.org/search/cs?searchtype=author&query=Hoffman%2C+S+C), [Stephanie Houde](https://arxiv.org/search/cs?searchtype=author&query=Houde%2C+S), [Kalapriya Kannan](https://arxiv.org/search/cs?searchtype=author&query=Kannan%2C+K), [Pranay Lohia](https://arxiv.org/search/cs?searchtype=author&query=Lohia%2C+P), [Jacquelyn Martino](https://arxiv.org/search/cs?searchtype=author&query=Martino%2C+J), [Sameep Mehta](https://arxiv.org/search/cs?searchtype=author&query=Mehta%2C+S), [Aleksandra Mojsilovic](https://arxiv.org/search/cs?searchtype=author&query=Mojsilovic%2C+A), [Seema Nagar](https://arxiv.org/search/cs?searchtype=author&query=Nagar%2C+S), [Karthikeyan Natesan Ramamurthy](https://arxiv.org/search/cs?searchtype=author&query=Ramamurthy%2C+K+N), [John Richards](https://arxiv.org/search/cs?searchtype=author&query=Richards%2C+J), [Diptikalyan Saha](https://arxiv.org/search/cs?searchtype=author&query=Saha%2C+D), [Prasanna Sattigeri](https://arxiv.org/search/cs?searchtype=author&query=Sattigeri%2C+P), [Moninder Singh](https://arxiv.org/search/cs?searchtype=author&query=Singh%2C+M), [Kush R. Varshney](https://arxiv.org/search/cs?searchtype=author&query=Varshney%2C+K+R), [Yunfeng Zhang](https://arxiv.org/search/cs?searchtype=author&query=Zhang%2C+Y). October 3rd 2018. [**https://arxiv.org/abs/1810.01943**](https://arxiv.org/abs/1810.01943)

**What:** This paper introduces a new open source Python toolkit for algorithmic fairness

**How:** The main objectives of this toolkit are to help facilitate the transition of fairness research algorithms to use in an industrial setting and to provide a common framework for fairness researchers to share and evaluate algorithms.

# **Algorithmic decision making and the cost of fairness**

# **Citation:** [Sam Corbett-Davies](https://arxiv.org/search/cs?searchtype=author&query=Corbett-Davies%2C+S), [Emma Pierson](https://arxiv.org/search/cs?searchtype=author&query=Pierson%2C+E), [Avi Feller](https://arxiv.org/search/cs?searchtype=author&query=Feller%2C+A), [Sharad Goel](https://arxiv.org/search/cs?searchtype=author&query=Goel%2C+S), [Aziz Huq](https://arxiv.org/search/cs?searchtype=author&query=Huq%2C+A), June 10th 2017 <https://arxiv.org/abs/1701.08230>

**What:** Algorithms are regularly used to decide whether defendants awaiting trial are too dangerous to be released. In some cases, black defendants are substantially more likely than white defendants to be incorrectly classified as high risk. To mitigate such disparities, several techniques have recently been proposed to achieve algorithmic fairness.

**How:** In this paper the authors reformulate algorithmic fairness as constrained optimization: the objective is to maximize public safety while satisfying formal fairness constraints designed to reduce racial disparities.

**Conclusion:** By analyzing data from Broward County, the authors find that optimizing for public safety yields stark racial disparities; conversely, satisfying past fairness definitions means releasing more high-risk defendants, adversely affecting public safety. And algorithms have the potential to improve the efficiency and equity of decisions, but their design and application raise complex questions for researchers and policymakers.

# **Decoupled classifiers for fair and efficient machine learning**

# **Citation:** [Cynthia Dwork](https://arxiv.org/search/cs?searchtype=author&query=Dwork%2C+C), [Nicole Immorlica](https://arxiv.org/search/cs?searchtype=author&query=Immorlica%2C+N), [Adam Tauman Kalai](https://arxiv.org/search/cs?searchtype=author&query=Kalai%2C+A+T), [Max Leiserson](https://arxiv.org/search/cs?searchtype=author&query=Leiserson%2C+M), July 21, 2017 https://arxiv.org/pdf/1707.06613v1.pdf

**What:** The paper considers how to use a sensitive attribute such as gender or race to maximize fairness and accuracy, assuming it is legal and ethical.

**How:** In this paper they explore decoupled classification systems, in which a separate classifier is trained on each group.

**Conclusion:** Experiments demonstrate that decoupling can reduce the loss on some datasets for some potentially sensitive features

**Evidence and Actions on Mortgage Market Disparities: Research, Fair lending Enforcement and Consumer Protection**

# **Citation:** Marsha J. Courchane & Stephen L. Ross, 2018. "[Evidence and Actions on Mortgage Market Disparities: Research, Fair Lending Enforcement and Consumer Protection](https://ideas.repec.org/p/hka/wpaper/2018-052.html)," [Working Papers](https://ideas.repec.org/s/hka/wpaper.html) 2018-052, Human Capital and Economic Opportunity Working Group. http://web2.uconn.edu/economics/working/2018-14.pdf

**What:** The authors present overviews of the research on discrimination in mortgage underwriting and pricing, the experiences of minority borrowers prior to and during the financial crisis, as well as discuss the history of legal cases alleging disparate treatment of minority borrowers.

**How:** By using the existing research and legal discussions as a background, they discuss and examine mortgage regulations as well as recent developments in the FinTech industry including machine learning.

**Conclusion:** The empirical evidence is clear that unexplained racial and ethnic differences in both mortgage underwriting and prices are relatively small when looking within lender and controlling for product attributes. Yet, race and ethnicity appear to play a role in determining mortgage market outcomes in the U.S. economy. The authors believe many factors, including information, attachment to financial markets, pre- application assistance levels and even shopping behavior contributed to the long-standing failure of prime qualified minority borrowers to access the conventional, conforming loan market.

# 

# **Peer-to-peer lending and bias in crowd decision-making**

**Citation:** Singh P, Uparna J, Karampourniotis P, Horvat E-A, Szymanski B, Korniss G, et al. (2018) Peer-to-peer lending and bias in crowd decision-making. PLoS ONE 13(3): e0193007. <https://doi.org/10.1371/journal.pone.0193007>

**What:** The authors investigate the “flat-world” hypothesis, which is the idea that globalization eventually leads to economic equality by analyzing crowdfinancing data, to see if it creates opportunities for the world’s poor.

**How:**The authors analyzed 600 000 peer-to-peer loans made to individual lenders in more than 220 countries to borrowers in 80 countries. They used regression analysis to predict bias in country-pair transactions based on variables such as GDP, geographical distance and so forth, checked for deviations of the co-country network of loans and checked the potential susceptibility of the network to shocks that could change the system’s ability potential for flatness.

**Conclusion:** The authors found continued and increased bias in an inter-country, peer-to-peer crowdfinancing network. The biases are reinforced and made stronger by the rapid growth of the platform itself (“rich gets richer” effect). However, by removing a few high-volume lenders or high-transaction links could cause the network’s flatness to increase significantly. In this way, the bias is directly linked with the dominance of a few big players.

**Other potential resources:**

Herding-effect:

<https://www.sciencedirect.com/science/article/pii/S1094996810000435>

<https://dl.acm.org/citation.cfm?id=1963483>

<https://www.sciencedirect.com/science/article/pii/S1567422312000154>

Gender / residency / profiles / biases:

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1738837>

<http://dspace.unive.it/bitstream/handle/10579/12866/989061-1229955.pdf?sequence=2>

<https://business.depaul.edu/about/centers-institutes/financial-services/Documents/1_3_Hasan.pdf>

<https://link.springer.com/article/10.1007/s10796-017-9751-5>

<https://www.tandfonline.com/doi/abs/10.1080/00036846.2016.1262526>

<https://www.sciencedirect.com/science/article/pii/S0148619518300390>

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2271298>

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3207163>

Language use / appearance:

<https://nca.tandfonline.com/doi/abs/10.1080/00909882.2010.536844#.W_hhaC2DrmE>

<https://academic.oup.com/rfs/article-abstract/25/8/2455/1570804>

Factors for loan defaults:

<https://www.tandfonline.com/doi/abs/10.1080/00036846.2014.962222>

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3178461>

<https://www.sciencedirect.com/science/article/pii/S0167718716302776>

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0139427#>

Machine learning models:

<https://www.wiseleywu.com/loan_default/final_report.pdf>

<http://cs229.stanford.edu/proj2017/final-reports/5228410.pdf>

<http://cs229.stanford.edu/proj2015/199_report.pdf>

<https://pdfs.semanticscholar.org/f33e/8956e43a195df522a9877deb4b44f15703f7.pdf>

<https://www.sciencedirect.com/science/article/pii/S1567422317300996>

<https://www.lexjansen.com/sesug/2016/BF-236_Final_PDF.pdf>