UNFAIRNESS BY ALGORITHM:

DISTILLING THE HARMS OF AUTOMATED DECISION-MAKING

December 2017

Overview

Analysis of personal data can be used to improve services, advance research, and combat discrimination. However, such analysis can also create valid concerns about differential treatment of individuals or harmful impacts on vulnerable communities. These concerns can be amplified when automated decision-making uses sensitive data (such as race, gender, or familial status), impacts protected classes, or affects individuals' eligibility for housing, employment, or other core services. When seeking to identify harms, it is important to appreciate the context of interactions between individuals, companies, and governments—including the benefits provided by automated decision-making frameworks, and the fallibility of human decision-making.

Recent discussions have highlighted legal and ethical issues raised by the use of sensitive data for hiring, policing, benefits determinations, marketing, and other purposes. These conversations can become mired in definitional challenges that make progress towards solutions difficult. There are few easy ways to navigate these issues, but if stakeholders hold frank discussions, we can do more to promote fairness, encourage responsible data use, and combat discrimination.

To facilitate these discussions, the Future of Privacy Forum (FPF) attempted to identify, articulate, and categorize the types of harm that may result from automated decision-making. To inform this effort, FPF reviewed leading books, articles, and advocacy pieces on the topic of algorithmic discrimination. We distilled both the harms and potential mitigation strategies identified in the literature into two charts. We hope you will suggest revisions, identify challenges, and help improve the document by contacting lsmith@fpf.org. In addition to presenting this document for consideration for the FTC Informational Injury workshop, we anticipate it will be useful in assessing fairness, transparency and accountability for artificial intelligence, as well as methodologies to assess impacts on rights and freedoms under the EU General Data Protection Regulation.

The Chart of Potential Harms from Automated Decision-Making

This chart groups the harms identified in the literature into four broad "buckets"—loss of opportunity, economic loss, social detriment, and loss of liberty—to depict the various spheres of life where automated decision-making can cause injury. It also notes whether each harm manifests for individuals or collectives, and as illegal or simply unfair.

We hope that by identifying and categorizing the harms, we can begin a process that will empower those seeking solutions to mitigate these harms. We believe that a more clear articulation of harms will help focus attention and energy on potential mitigation strategies that can reduce the risks of algorithmic discrimination. We attempted to include all harms articulated in the literature in this chart; we do not presume to establish which harms pose greater or lesser risks to individuals or society.

The Chart of Potential Mitigation Sets

This chart uses FPF's taxonomy to further categorize harms into groups that are sufficiently similar to each other that they could be amenable to the same mitigation strategies.

Attempts to solve or prevent this broad swath of harms will require a range of tools and perspectives. Such attempts benefit by further categorization of the identified harms, into five groups of similar harms. These groups include: (1) individual harms that are illegal; (2) individual harms that are simply unfair, but have a corresponding illegal analog; (3) collective/societal harms that have a corresponding individual illegal analog; (4) individual harms that are unfair and lack a corresponding illegal analog; and (5) collective/societal harms that lack a corresponding individual illegal analog. The chart includes a description of the mitigation strategies that are best positioned to address each group of harms.

There is ample debate about whether the lawful decisions included in this chart are fair, unfair, ethical, or unethical. Absent societal consensus, these harms may not be ripe for legal remedies.

Potential Harms from Automated Decision-Making

Individual Harms		Collective / Societal Harms
Illegal	**Unfair**	
Loss of Opportunity		
Employment Discrimination E.g. Filtering job candidates by race or genetic/health information	E.g. Filtering candidates by work proximity leads to excluding minorities	**Differential Access to Job Opportunities**
Insurance & Social Benefit Discrimination E.g. Higher termination rate for benefit eligibility by religious group	E.g. Increasing auto insurance prices for night-shift workers	**Differential Access to Insurance & Benefits**
Housing Discrimination E.g. Landlord relies on search results suggesting criminal history by race	E.g. Matching algorithm less likely to provide suitable housing for minorities	**Differential Access to Housing**
Education Discrimination E.g. Denial of opportunity for a student in a certain ability category	E.g. Presenting only ads on for-profit colleges to low-income individuals	**Differential Access to Education**
Economic Loss		
Credit Discrimination E.g. Denying credit to all residents in specified neighborhoods ("redlining")	E.g. Not presenting certain credit offers to members of certain groups	**Differential Access to Credit**
Differential Pricing of Goods and Services E.g. Raising online prices based on membership in a protected class	E.g. Presenting product discounts based on "ethnic affinity"	**Differential Access to Goods and Services**
	Narrowing of Choice E.g. Presenting ads based solely on past "clicks"	**Narrowing of Choice for Groups**
Social Detriment		
	Network Bubbles E.g. Varied exposure to opportunity or evaluation based on "who you know"	**Filter Bubbles** E.g. Algorithms that promote only familiar news and information
	Dignitary Harms E.g. Emotional distress due to bias or a decision based on incorrect data	**Stereotype Reinforcement** E.g. Assumption that computed decisions are inherently unbiased
	Constraints of Bias E.g. Constrained conceptions of career prospects based on search results	**Confirmation Bias** E.g. All-male image search results for "CEO," all-female results for "teacher"
Loss of Liberty		
	Constraints of Suspicion E.g. Emotional, dignitary, and social impacts of increased surveillance	**Increased Surveillance** E.g. Use of "predictive policing" to police minority neighborhoods more
Individual Incarceration E.g. Use of "recidivism scores" to determine prison sentence length (legal status uncertain)		**Disproportionate Incarceration** E.g. Incarceration of groups at higher rates based on historic policing data

Potential Mitigation Sets

Harms	Description	Mitigation Tools
Individual Harms – Illegal		
Employment Discrimination **Insurance & Social Benefit Discrimination** **Housing Discrimination** **Education Discrimination** **Credit Discrimination** **Differential Pricing** **Individual Incarceration**	Existing law defines impermissible outcomes, often specifically for protected classes	• **Data methods** to ensure proxies are not used for protected classes & data does not amplify historical bias • **Algorithmic design** to carefully consider whether to use protected status inputs & trigger manual reviews • **Laws & policies** that use data to identify discrimination
Individual Harms – Unfair (with illegal analog)		
Employment Discrimination **Insurance & Social Benefit Discrimination** **Housing Discrimination** **Education Discrimination** **Credit Discrimination** **Differential Pricing** **Individual Incarceration**	Individual harms that could be considered illegal if they involved protected classes, but do not in this case	• **Business processes** to index concerns; ethical frameworks & best practices to monitor & evaluate outcomes • **Laws & policies** include tools like DPIAs to measure impact or enable rights to explanation
Collective/Societal Harms (with illegal analog)		
Differential Access to Job Opportunities **Differential Access to Insurance Benefits** **Differential Access to Housing** **Differential Access to Education** **Differential Access to Credit** **Differential Access to Goods & Services** **Disproportionate Incarceration**	Group level impacts that are not legally prohibited, though related individual impacts could be illegal	• Same as above section • **Laws & policies** should consider offline analogies & whether it is appropriate for industry to identify & mitigate
Individual Harms – Unfair (without illegal analog)		
Narrowing of Choice **Network Bubbles** **Dignitary Harms** **Constraints of Bias** **Constraints of Suspicion**	Individual impacts for which we do not have legal rules. Mitigation may be difficult or undesirable absent a defined set of societal norms	• **Business processes** to index concerns, ethical frameworks & best practices to monitor & evaluate outcomes • **Laws & policies** should consider whether it is appropriate to expect industry to identify & enforce norms
Collective/Societal Harms (without illegal analog)		
Narrowing of Choice for Groups **Filter Bubbles** **Stereotype Reinforcement** **Confirmation Bias** **Increased Surveillance of Groups**	Group level impacts for which we do not have legal rules or societal agreement as to what constitutes a harm	• Same as above section
Key		
Loss of Opportunity	**Economic Loss**	**Social Stigmatization** / **Loss of Liberty**

Working Definitions: Harms

Automated Decision: The direct output or indirect result from an automated program analyzing individual or aggregate data. This includes pre-programmed algorithms and those that evolve via machine learning techniques.

Illegal: Examples in this category represent harms that are illegal under several U.S. civil rights laws, which generally protect core classifications—such as race, gender, age, and ability—against discrimination, disparate treatment, and disparate impact.

Unfair: Examples in this category represent actions that are typically legal, but nonetheless trigger notions of unfairness. Like the "illegal" category, some examples here may be differently classified depending on the legal regime.

Collective / Societal Harms: This category represents overall negative effects to society that are chiefly collective, rather than individual in nature.

Loss of Opportunity: This group broadly describes harms occurring within the domains of the workplace, housing, social support systems, healthcare, and education.

Economic Loss: This group broadly describes harms that primarily cause financial injury or discrimination in the marketplace for goods and services.

Social Detriment: This group broadly describes harms to one's sense of self, self worth, or community standing relative to others.

Loss of Liberty: This group broadly describes harms that constrain one's physical freedom and autonomy.

Working Definitions: Mitigation

Individual Harms – Illegal: The harms in this category are those for which American law defines outcomes that are not legally permissible. These harms typically become legally cognizable because they impact legally protected classes in a manner that is defined as impermissible under existing law. Notably, disparate impact may be relevant to illegality regardless of intent in some areas.

Individual Harms – Unfair (with illegal analog): The individual harms in this category do not involve protected classes, but could be considered illegal if protected classes were implicated. For example, while price discrimination based on race could be illegal under the Fair Credit Reporting Act or Civil Rights Act, price discrimination based on computer operating system of the user is not protected under the law. Nonetheless, automated decision-making enables a growing number of personalized distinctions. Some may consider these distinctions unfair or unethical.

Collective/Societal Harms (with illegal analog): In this category, impacts at the group level may not be legally prohibited, but individual impacts could be illegal under different circumstances. While rules may prohibit disparate treatment of protected classes, differential treatment of groups that are not legally protected may not be considered illegal. For example, systematically failing to hire people of a certain race may be illegal, but systematically failing to hire Apple computer users or Red Sox fans is not protected under the law, though some may consider it unfair.

Individual Harms – Unfair (without illegal analog): This category applies to impacts on individuals for which we do not have legal rules. Some, such as narrowing of choice and network bubbles, may be harms that are newly enabled by the growth of technology platforms. Others, such as the the constraints of bias or the constraints of suspicion, have been challenges in the analog world for decades.

Collective/Societal Harms (without illegal analog): This category includes collective outcomes for which we do not have legal rules. As with the prior group, some of these harms—such as narrowing of choice for groups and filter bubbles—have become more frequent due to increased reliance on algorithmic personalization techniques. Stereotype reinforcement is as old as time, but can be compounded by the volume of information available online. Confirmation bias and increased surveillance of groups have been challenges in society for decades, if not since its inception.

Reviewed Literature

The alphabetized list below captures the literature FPF has reviewed to date for this effort. We welcome suggestions for further materials to review to lsmith@fpf.org.

- Aaron Reike, *Don't let the hype over "social media scores" distract you*, EQUAL FUTURE (2016).
- Alessandro Acquisti & Christina Fong, *An Experiment in Hiring Discrimination via Online Social Network*, presented at Privacy Law Scholars Conference (2016).
- Alethea Lange et al., *A User-Centered Perspective on Algorithmic Personalization*, presented at the Fed. Trade Comm'n PrivacyCon Conference (2017).
- Allan King & Marko Mrkonich, *"Big Data" and the Risk of Employment Discrimination*, 68 OKLA. L. REV. 555 (2016).
- Andrew Tutt, *An FDA for Algorithms*, 67 ADMIN. L. REV. 1 (2016).
- Aniko Hannak et al., *Bias in Online Freelance Marketplaces: Evidence from TaskRabbit*, presented at the Workshop on Data and Algorithmic Transparency (Nov. 2016).
- CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION (2016).
- Christian Sandvig et al., Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms, presented at the Int'l Comm'cn Ass'n Conference on Data and Discrimination: Converting Critical Concerns into Productive Inquiry (2014).
- Daniel Solove, *A Taxonomy of Privacy*, 154 U. PENN. L. REV. 3 (2016).
- Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014).
- EXEC. OFF. OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES (2014).
- EXEC. OFF. OF THE PRESIDENT, BIG DATA: A REPORT ON ALGORITHMIC SYSTEMS, OPPORTUNITY, AND CIVIL RIGHTS (2016).
- FEDERAL TRADE COMMISSION, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? (Jan. 2016).
- Frank Pasquale & Danielle Keats Citron, *Promoting Innovation While Preventing Discrimination: Policy Goals for the Scored Society*, 89 WASH. L. REV. 1413 (2014).
- Jennifer Valentino-Devries, Jeremy Singer-Vine, Ashkan Soltani, *Websites Vary Prices, Deals Based on Users' Information*, WALL ST. J. (Dec. 24, 2012).
- Joshua Kroll et al., *Accountable Algorithms*, 165 U. PENN. L. REV. 633 (2016).
- Juhi Kulshrestha et al., *Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media*, presented at the Workshop on Data and Algorithmic Transparency (2016).
- Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C.L. REV. 93 (2014).
- Latanya Sweeney, *Discrimination in Online Ad Delivery*, COMMC'NS OF THE ASS'N OF COMPUTING MACHINERY (2013).
- Lee Rainie & Jana Anderson, *Code-Dependent: Pros and Cons of the Algorithm Age*, PEW RESEARCH CENTER (2017).
- Mark MacCarthy, *Student Privacy: Harm and Context*, 21 INT'L REV. OF INFO. ETHICS 11 (2014).
- Mary Madden, Michele Gilman, Karen Levy & Alice Marwick, *Privacy , Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans*, Wash. U. L. Rev __ (forthcoming) (Mar. 2017).
- Megan Garcia, *How to Keep Your AI From Turning Into a Racist Monster*, WIRED (2017).
- Moritz Hardt, Eric Price & Nathan Srebro, *Equality of Opportunity in Supervised Learning*, presented at the Conference on Neural Info. Processing Sys. (2016).
- Motahhare Eslami et al., *Reasoning about Invisible Algorithms in the News Feed*, presented at the Ass'n of Computing Machinery Special Interest Gp. on Computer-Human Interaction (2015).
- Muhammad Zafar et al., *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment*, presented at the Int'l World Wide Web Conference (2017).
- Nanette Byrnes, *Why We Should Expect Algorithms to be Biased*, MIT TECHNOLOGY REVIEW (2016).
- NEW AMERICA & OPEN TECH. INST., DATA AND DISCRIMINATION: COLLECTED ESSAYS (S.P. Gangadharan, Ed. 2014).
- Omer Tene and Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 Nw. J. Tech. & Intell. Prop.239 (2013).
- PAM DIXON & ROBERT GELLMAN, THE SCORING OF AMERICA: HOW SECRET CONSUMER SCORES THREATEN YOUR PRIVACY AND YOUR FUTURE, WORLD PRIVACY FORUM (2014).
- Pauline Kim, *Data-Driven Discrimination at Work*, 59 WILLIAM & MARY L. REV. ___ (2017).
- Peter Swire, *Lessons From Fair Lending Law for Fair Marketing and Big Data* (2014)
- PROPUBLICA, Machine Bias Investigative Series, https://www.propublica.org/series/machine-bias
- Sandra Wachter, Brent Mittelstadt, & Luciano Floridi, *Why a right to explanation of automated decision making does not exist in the General Data Protection Regulation* (2016).
- Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016).
- UPTURN, CIVIL RIGHTS, BIG DATA, AND OUR ALGORITHMIC FUTURE (2014).

1400 EYE STREET, NW | SUITE 450 | WASHINGTON, DC 20005 • FPF.ORG

202.768.8950

INFO@FPF.ORG | MEDIA@FPF.ORG