Definition of Algorithmic Unfairness

Authors
Allison Woodruff (Security & Privacy UX)
Andy Schou (Security & Privacy Nightwatch)

Last Updated
February 2017

go/algorithmic-unfairness-definition

Goals

Our goal is to create a company-wide definition of algorithmic unfairness that:

1. **Articulates the full range of algorithmic unfairness that can occur in products.** This definition should be robust across products and organizational functions.
2. **Establishes a shared understanding of algorithmic unfairness** for use in the development of measurement tools, product policy, incident response, and other internal functions.[1]
3. **Is broadly consistent with external usage of the concept.** While it is not a goal at this time to release this definition externally, it should represent external concerns so that we can ensure our internal functions address these concerns.[2,3,4]

Non-Goals

The following are not goals for this document:

1. **Specify whether and how Google will take action** on potential instances of unfairness that involve the use of an algorithm. This will fall instead to product policy.
2. **Describe the consequences of algorithmic unfairness and why they matter**.

Definition

"algorithmic unfairness" means unjust or prejudicial treatment of people that is related to sensitive characteristics such as race, income, sexual orientation, or gender,[5] through algorithmic systems or algorithmically aided decision-making.

---

[1] The discussion in Lipton (2016) suggests that instances that present as identical product behavior may upon investigation be revealed to have substantially different root causes and therefore require different remediations. Because the nature of an instance may not be evident *a priori*, a wide range of instances are likely to be reported and investigated together. (Zachary Chase Lipton. The Deception of Supervised Learning. KDnuggets News 16:n33, September 2016.)

[2] Peter Swire. Lessons from Fair Lending Law for Fair Marketing and Big Data. Future of Privacy Forum, and presented before the Federal Trade Commission Workshop on "Big Data: A Tool for Inclusion or Exclusion?" (2014).

[3] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. 104 California Law Review 671 (2016).

[4] ACM US Public Policy Council. Statement on Algorithmic Transparency and Accountability. January 12, 2017.

[5] This list is not intended to be exhaustive, and additional examples of sensitive characteristics appear in the FAQ.

Scope

The following are in scope for algorithmic unfairness:

In algorithms that drive **predictive** systems (e.g., personalization or device financing), this definition encompasses automated actions significantly adverse to the interests of a user or group of users,[6] on the basis of a characteristic that is sensitive in the context of a particular interaction.

In algorithms that drive **representational** systems (e.g., search), this definition encompasses the implied endorsement of content likely to shock, offend, or upset users sharing a sensitive characteristic, or to reinforce social biases.[7]

The following are *not* in scope for this definition but may be covered by other internal definitions and policies:

1. **Biased content**, for example user-generated content that appears in products (while the content itself is not in scope, the algorithmic handling of such content may be in scope)
2. **Insensitive designs** which occasion unfairness, and/or interface designs that are less usable by groups with sensitive characteristics
3. **Google's internal decisions** such as hiring or compensation, which may be influenced by algorithms

Test Cases for Scope

The following are illustrative examples of what is (and is not) in scope for algorithmic unfairness. Of the examples that are in scope, a subset may be determined by product policy to require remediation.

Google Display Ads for High-Paying Jobs

CMU published a study in 2015 with experiment-based observations, arguing that Google's ad serving system perpetuates gender bias on the basis of two campaigns that were found to target high salary jobs at male users on the Times of India website.[8] The effect was highly sensitive to the ads from this particular service, and the same effect was not reproduced in several other experiments by the same authors. The cause was found to be higher CPA (cost per conversion) for

[6] This includes effects which are small for a single instance but have a significant cumulative effect. As Greenwald et al. observe, statistically small effects can have substantial societal impact when they apply to many people, or if they apply repeatedly to the same person. (Anthony G. Greenwald, Mahzarin R. Banaji, and Brian A. Nosek. Statistically small effects of the Implicit Association Test can have societally large effects. Journal of Personality and Social Psychology, 108:553–561, 2015.)

[7] Note that even small reinforcement biases in systems can be magnified via positive feedback loops, since biased representations can influence human behavior which is in turn fed back into training data for those systems.

[8] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings. PETS 2015, pp. 92-112, June 2015.

female users for one campaign (notably with a higher CTR for female users) and advertiser targeting to male-only users for the other.[9]

In scope for algorithmic unfairness. In the first case, bias in user behavior trained the system to have biased targeting. In the second case, advertiser-selected criteria were related to a sensitive characteristic.

Facebook Computation of Unregulated FICO Scores

Researchers raised concerns in 2015 about Facebook computing non-regulated credit scores based on user activity.[10] One substantial issue they commented on is that credit scores are regulated by the Equal Credit Opportunity Act of 1974, which prohibits creditors from discriminating against applicants on the basis of race, religion, national origin, sex, marital status, age, or receiving public assistance; however, the non-regulated scores did not appear to have such restrictions.[11]

In scope for algorithmic unfairness. Based on the user's behavior, the system automatically computed sensitive characteristics that could adversely affect their financial opportunities.

Chumhum Suppression of Businesses in High-Crime Areas

An episode of the primetime television drama "The Good Wife" featured a tortious interference case against a fictional search engine company (Chumhum) for releasing a maps application that suppressed businesses in high crime areas.[12]

In scope for algorithmic unfairness. Content was excluded by an algorithm in a way that disproportionately affected people who owned businesses in neighborhoods correlated with the sensitive characteristics race and income.

West African Spam Filters

An external researcher poses the following: "One question is whether the design of spam filters could make certain individuals more susceptible to having their legitimate messages diverted to spam folders. For example, does being located in a hotbed of Internet fraud or spam activity, say West Africa (Nigeria or Ghana) or Eastern Europe, create a tendency for one's messages to be mislabeled as spam?"[13]

In scope for algorithmic unfairness. In this hypothetical, the system learns associations and downgrades or excludes content related to sensitive characteristics such as national origin and race.

---

[9] Giles Hogben, Alex McPhillips, Vinay Goel, and Allison Woodruff. Allegations of Algorithmic Bias: Investigation and Meta-Analysis. September 2016. go/allegations-of-algorithmic-bias

[10] Tressie McMillan Cottom. Credit Scores, Life Chances, and Algorithms. May 30, 2015.

[11] Astra Taylor and Jathan Sadowski. How Companies Turn Your Facebook Activity Into a Credit Score: Welcome to the Wild West of Data Collection Without Regulation. *The Nation*, May 27, 2015.

[12] The Good Wife, Episode "Discovery". CBS, first aired November 22, 2015.

[13] Jenna Burrell. How the machine "thinks": Understanding opacity in machine learning algorithms. Big Data & Society 3(1):1-12, 2016.

Google Photos Use of Gorillas Label

In June 2015, a web developer posted on Twitter that Google Photos had tagged an image showing him and a friend at a concert with the label "Gorillas". "Of all terms, of all the derogatory terms to use," Alciné said later, "that one came up." According to a post mortem, Google executives noticed Alciné's tweets within one hour. A Googler reached out through Twitter for permission to access the user's photos, and the issue was identified and resolved. In the short term, the Photos team stopped suggesting the "Gorillas" tile in the Explore page and stopped showing Search results for queries relevant to gorillas; in the long term, the team pledged to investigate the image annotation models that generate false positives for the gorilla label and work on the image annotation pipeline and quality evaluation processes.[14]

In scope for algorithmic unfairness. The system drew an incorrect inference that is offensive to members of a given race.

Autocomplete Results for Trayvon Martin

In 2013, Autocomplete results showed negative results for Travyon Martin (e.g., "drug dealer"), but more positive results for George Zimmerman (e.g., "hero").[15]

In scope for algorithmic unfairness. The system processed and showed offensive content from users, in a way that could be seen to reinforce existing social biases regarding race.

Google Search Results for Black Girls & Pornography

In 2013, a researcher expressed concern that search queries on Google for "black girls" historically yielded a high percentage of pornography.[16]

In scope for algorithmic unfairness. The system showed results that are offensive and reinforce existing social biases regarding race.

Google Image Search Results for Physicists

Google image search results for "physicist" show predominantly men. In reality, roughly 20% of physicists are women. First, imagine that the image search results show 1% women. Second, imagine instead that the image search results show 20% women.

In scope for algorithmic unfairness. In the first case, when the image search results show only 1% when the reality is 20%, the system is *amplifying* an existing bias in society. In the second case, when the image search results show a percentage similar to the current reality, the system is *reflecting* an

---

[14] Giles Hogben, Alex McPhillips, Vinay Goel, and Allison Woodruff. Allegations of Algorithmic Bias: Investigation and Meta-Analysis. September 2016. go/allegations-of-algorithmic-bias

[15] Safiya Umoja Noble. Trayvon, Race, Media and the Politics of Spectacle.The Black Scholar. 44(1):12-29, Spring 2014.

[16] Safiya Umoja Noble. Google Search: Hyper-visibility as a Means of Rendering Black Women and Girls Invisible. InVisible Culture: Issue 19, October 29, 2013.

existing bias in society. Both cases fall in scope for algorithmic unfairness because they reinforce a stereotype about the role of women in a scientific discipline. However, while both are in scope for algorithmic unfairness, remediation may be more likely in the former case. (As with the other cases, whether and how to remediate would fall to product policy.)

Microsoft Kinect

News media in 2010 and 2017 reported on developers and users of Microsoft's Kinect experiencing difficulties with facial recognition or motion detection for users with dark skin in low light conditions. Affected users could not use certain system functions including automatic sign in to the user's profile,[17] or use software that required motion detection to function.[18]

In scope for algorithmic unfairness. The predictive system performed poorly for users with dark skin, which is associated with the sensitive characteristic race. Users considering purchasing the Kinect or software that requires the Kinect would be unlikely to suspect that their skin color might make it difficult or impossible to use advertised features (such as automatic login using facial recognition, gestural menu interactions, motion controlled gaming). Gaming hardware and software purchases are commonly non-refundable after being opened, so a user who realized the Kinect or its software could not track them would be left with a non-functional product they could not return.

Survivalist Game on the Play Store

In January of 2016, a series of Twitter posts and a Change.org Petition raised concerns about a Survivalist game on the Play Store, claiming it gamified killing aborigines.

Not in scope for algorithmic unfairness. There were concerns that the content was offensive, but an algorithm was not involved.

Airbnb Acceptance Rates

Researchers from the Harvard Business School conducted an experiment on Airbnb and found that applications from guests with distinctively African-American names were less likely to be accepted relative to identical guests with distinctively white names. The study could not identify whether the unfairness was based on race, socioeconomic status, or some other factor.[19]

Not in scope for algorithmic unfairness. The unfairness was on the part of individual users (“sellers” on Airbnb) and there was no algorithmic component to the unfairness. While this is out of scope for algorithmic unfairness, Airbnb did however recognize that their design choices unrelated to algorithms could be facilitating unfair behavior. Accordingly, it pursued non-algorithmic remediations, such as modifying its user interface design and its end user policy.[20]

---

[17] Brendan Sinclair. Kinect has problems recognizing dark-skinned users? Gamespot, 2010.

[18] Andy Trowers. How We Accidentally Made a Racist Videogame. Kotaku, 2017.

[19] Benjamin Edelman, Michael Luca, and Dan Svirsky. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. Harvard Business School NOM Unit Working Paper (16-069), 2015.

[20] Laura W. Murphy. Airbnb’s Work to Fight Discrimination and Build Inclusion: A Report Submitted to Airbnb. September 2016.

FAQ

Is personalization the same as algorithmic unfairness?
Generally, no. Personalization is algorithmic behavior that presents different results to different users. Personalization is very often beneficial to users and is not necessarily unfair. In fact, personalization may sometimes remediate unfairness by identifying an individual's interests more precisely than would the use of a sensitive characteristic.[21] However, in some cases personalization may be strongly associated with a sensitive characteristic in a way that causes significant harm to a user or users, in which case it would be algorithmic unfairness.

Can any characteristic be sensitive?
No. For example, "people who like yellow" or "people with pets" are unlikely to be sensitive.[22]

Which characteristics are sensitive?
Sensitive characteristics may be determined by legal considerations or by more broad principles. Legally determined characteristics may vary by jurisdiction, sector, or other factors. Sensitive characteristics determined by more broad principles are especially likely to include characteristics that are associated with less privileged or marginalized populations (particularly when such characteristics are immutable), may be socially undesirable, or may be associated with civil liberties.

Specific examples may include race/ethnic origin, gender identity, sexual orientation, religion, political party, disability, age, nationality, veteran status, socioeconomic status (including caste and homelessness), and immigrant status (including refugee and asylum seeking status). Further, there may be emergent sensitive characteristics for which we do not yet have a name, or which we have not yet anticipated.[23]

Are proxies for sensitive characteristics covered?
Yes. Proxies are characteristics that are highly correlated with a sensitive characteristic. Actions based on close proxies for sensitive characteristics are in scope for further investigation. For example, if interest in hip-hop music is highly correlated with being Black, targeting users who like hip hop music may result in algorithmic unfairness based on race, regardless of intent.

If a system's behavior is caused by societal bias, can it still be algorithmic unfairness?
Yes. Societal bias that is reflected in algorithmic behavior is a central issue in the external public, media, and regulatory concerns about algorithmic unfairness.[24,25,26,27] Additionally, one often doesn't

---

[21] James C. Cooper. Separation and Pooling. George Mason Law & Economics Research Paper No. 15-32, March 2016.
[22] http://civilrights.findlaw.com/civil-rights-overview/what-is-discrimination.html
[23] danah boyd, Karen Levy & Alice Marwick, The Networked Nature of Algorithmic Discrimination. In Seeta Peña Gangadharan, Virginia Eubanks, and Solon Barocas (eds), Data and Discrimination: Collected Essays. Washington, D.C.: Open Technology Institute, New America Foundation, pp. 53-57, 2014.
[24] Tarleton Gillespie. The Relevance of Algorithms. In Media Technologies: Essays on Communication, Materiality, and Society, edited by T. Gillespie, P. Boczkowski, and K. Foot. MIT Press, Cambridge, MA, 2012.
[25] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. 104 California Law Review 671 (2016).
[26] Executive Office of the President. Big Data: Seizing Opportunities, Preserving Values. May 2014.

know the root cause of algorithmic unfairness without an investigation, so any real-world process (e.g., incident response) is best served by encompassing a wide range of potential causes. Accordingly, this definition covers a wide range of sources, from machine learning classification errors to societal bias. However, the nature of the root cause may affect product policy's position on whether or how a given algorithmic behavior should be addressed.

If a representation is factually accurate, can it still be algorithmic unfairness?
Yes. For example, imagine that a Google image query for "CEOs" shows predominantly men. Even if it were a factually accurate representation of the world, it would be algorithmic unfairness because it would reinforce a stereotype about the role of women in leadership positions. However, factual accuracy may affect product policy's position on whether or how it should be addressed. In some cases, it may be appropriate to take no action if the system accurately affects current reality, while in other cases it may be desirable to consider how we might help society reach a more fair and equitable state, via either product intervention or broader corporate social responsibility efforts.

If a system's behavior is not intended, can it still be algorithmic unfairness?
Yes. If the behavior is unfair, it meets the definition regardless of the root cause.

If unfairness is executed by an algorithm but is the result of a human decision, can it still be algorithmic unfairness?
Yes. For example, if a human chooses unfair keywords for ads and those choices result in unfair algorithmic choices of what ads to show to a user or users, that would fall within the scope of algorithmic unfairness.

Does this definition include the use of data by multiple parties?
Yes. For instance, decisions which may result in data being provided to third parties via APIs, being sold to third parties, or acquired via federated identity should consider the possibility that that data could be used to power unfair algorithmic decision-making.

What is the relationship between this definition of algorithmic unfairness and the fairness measure in Hardt et al. (2016)?
The fairness measure in Hardt et al. (2016)[28] is a statistical guarantee that, in general, members of one category are classified with the same accuracy as members of another category. For example, if Black people who apply for credit cards are classified with lower accuracy than white people (e.g., Black people who would actually repay their loans are wrongly classified as bad credit risks), the fairness measure in Hardt et al. (2016) would find that to be a problem. However, if most people who happen to have a particular sensitive characteristic are correctly classified as being very likely to default on their credit cards (regardless of whether that classification is done based on that sensitive characteristic or based on some other associated variable), the Hardt et al. (2016) measure would not detect it as an issue. Therefore, the fairness measure in Hardt et al. (2016) statistically tests for a certain type of machine learning classification error that would constitute a specific type of algorithmic unfairness, encompassing some but not all of the cases covered by the current definition.

[27] ACM US Public Policy Council. Statement on Algorithmic Transparency and Accountability. January 12, 2017.
[28] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. arXiv.org, October 2016.

Acknowledgments

We thank the following for valuable discussions and input on this definition: Blaise Aguera y Arcas, Sarah Fox, Jen Gennai, Sarah Holland, Giles Hogben, Brad Krueger, Josh Lovejoy, Jakara Mato, Alex McPhillips, M. Mitchell, Divya Tam, and Jeff Warshaw.