How does bias creep in in algorithms?

Algorithms are mostly considered better as in more accurate, efficient and less biased as compared to humans. However, as with humans, there are limits to how much they can be trusted. Due to the perception that they are better and more accurate, they garner more trust than what they may deserve.

“Implicit signals also tell users how much they should trust a digital decision. For example, users may be more likely to trust a recidivism scoring tool that uses a numerical scale to rate defendants’ risk and is marketed for use by judges in sentencing hearings than they are to trust an online quiz called “how likely are you to end up in jail?” even though both may have the same research behind them and deliver similarly accurate results. When a decision-making system’s usefulness, accuracy, or underlying logic is misaligned with users’ expectations, it can induce misplaced trust resulting in harm.” 1

Most Prevalent Problem- direct application of historical data

How we select training dataset can lead to very disparate results.

Most ML algos identify trends with the basis being statistical correlations in the data. Which means, future prediction is heavily dependent on past data- or the past data that is even available. This can usually lead to discriminatory or unequal results due to the past influencing the model. “For example, if training data for an employment eligibility algorithm consists only of all past hires for a company, no matter how the target variable is defined, the algorithms may reproduce past prejudice, defeating efforts to diversify by race, gender, educational background, skills, or other characteristics.” 1

Unintentional Bias due to data collection where understanding of certain minorities is not complete

At times, data is collected under biased conditions – “for a purpose unrelated to the goal of the algorithm”1. Hence, this data fails to accurately represent the population forming skewness in the available data. We can look into the example of the smartphone app StreetBump released by the city of Boston to automatically report potholes. However, the point that was missed over ere was that the low income neighbourhoods have a lower likelihood of having smartphones which made the data collection method unintentionally disparate as there was comparatively little data from the lower income neighbourhoods.

In such cases it is necessary that we carefully audit data collection methods to make sure that other factors aren’t affecting the results of the data collection and equal representation of population is reflected in our data sample.

Bias in Digital Decisions

“Digital decisions may be harmful because (1) they are unreliable, (2) they are low-quality, (3) they cause disparate impacts among different groups or populations, (4) they are unaccountable, or some combination thereof.”1

The algorithms that are used for digital decisions are like a black box where there is very little transparency. Also, the quality of algorithms for digital decisions depends on what the consequences are for that decision. A very popular example is that of tampon ads, a poor advertising algo would be one tat advertises this to males. However, this is not particularly harmful when compared to a recidivism risk scoring algorithm which could cause life changing effects to people against whom the system is biased. For example, today the recidivism risk scoring is biased against blacks which are currently even used to decide and advice when a criminal should be released, what bond amounts should they be assigned, all these decisions which could be life changing.

The problem becomes highly serious when such algorithms are used for decisions related to economic opportunities, social services access, criminal justice results as here developers and decision makers need to more responsibility so that the algorithms make high quality decisions.

Importance of a good dataset in decision making system

As we discussed in the previous section, low quality data leads to low quality predictions. Hence, the inverse, a good data set is essential for high quality predictions. Examples of low quality datasets are – Too small, incomplete, skewed, or data not suited to the algorithm's purpose.

As we saw in the previous section, as ML algorithms learn to classify based on training data , issues in training data will be reflected in the decisions and affect their quality.

Another problem that result in low quality algorithm are prediction of rare or statistically difficult to predict events or qualities. Such type of events and things have a lot of different factors which are difficult to quantify or implement in a model or most of the times the data is very limited as compared to the population against whom it is used to predict. For eg, the prediction of terrorist attacks or violent crime. "This endeavour is so fraught with error and potentially meaningless statistical relationships that the hypothetical benefits may not be worth the downsides."1

Disparate Impacts due to biased decision systems –

Algorithms by themselves aren't objectively inaccurate or at times. "In fact, some of the most harmful algorithms are those that rely on statistically significant ways to sort people—often unintentionally—so that historically marginalized groups are denied opportunities or adversely targeted at higher rates."

As we have seen, historically, women and people of color have mostly been affected by discriminatory employment and have been under-represented at work. Present data is bound to have this bias and any algorithm based on it will favour white males.

- case in point google research paper- https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Current Solutions and why they fail

"Eliminating sensitive characteristics, such as race and gender, from an algorithm's equation does not solve the problem of disparate impacts. Predictive models can rely on variables or features that strongly correlate with race, gender, sexual orientation, age, or income. These are known as "proxies." For example, at least one recidivism risk assessment tool asks whether arrestees have a parent who has been imprisoned. However, this factor is also correlated with race, since African-Americans are arrested and imprisoned at higher rates than whites. Since African-Americans are statistically more likely to answer "yes" to this question, an algorithm relying on it could disproportionately rate African-American arrestees as more likely to commit future crimes."

Why are biased Algorithms harmful-

Such algorithms can be very harmful when used in relation to credit scoring, insurance, employment and governments programs where affected decisions are life changing. “Algorithms with systematic disparate impacts could violate antidiscrimination laws, such as the Fair Housing Act or the Equal Employment Opportunity Act.” Designers and algorithm developers must be particularly careful and have an ethical obligation to make sure that they are aware of the biases and taking those into account so that they don’t result in marginalising certain communities.

Unaccountable decisions and issues in reviewing them

Automated decision-making systems are often referred to as “black boxes,” because their existence and logic are unknown to the people whose lives they affect. Opaque decision-making systems can serve predatory loan advertisements or make sentencing decisions without the target ever knowing precisely why. We often don’t even know an algorithm was used to make a decision about us. This makes digital decisions difficult to review or challenge—there is often no way to inform an institution or seek redress when we suspect the algorithm got it wrong.

Lack of redress for unfair or incorrect automated decisions can lead to frustration, loss of control, and distrust of institutions. The most vulnerable among us may be the most likely to give up on opaque automated decisions that don’t work for them and to avoid decision-making institutions. Unreliable or unfair decisions that go unchallenged can contribute to bad feedback loops, which can make algorithms even more likely to marginalize vulnerable populations. When automated systems do not allow for user feedback, they create blind spots that prevent them from learning from their own bad decisions.

So What do we do? Solutions which can be put in practice

Academics, journalists, advocates, technologists, industry groups, and even the Obama Administration have raised concerns about the potential harms of unchecked, biased automation. Many of those same groups have also undertaken efforts to identify and rally support behind a set of principles that describe shared values in response to the problem.

Current cultural perceptions of automated decision making technology are out of step with the technical reality.

Data does capture facts but it’s not by default! (hence, not completely objective)

Many present-day values and beliefs about technology implicitly or explicitly endorse public trust in data-driven decision making as objective or more fair than human decisionmaking. For example, the belief that data captures a neutral or objective view on reality is widespread, and evoked often to explain otherwise improper conclusions like gender playing a role in creditworthiness. And of course data often does capture facts and provide an objective (or at least measurable) insight, but that does not happen by default.

Responsible use of data is a necessity. Disrupting beliefs like ‘data is objective’ requires replacing misconceptions about technology with principles that are ethically and technically sound. The most effective statements recognize the technical pitfalls of relying on big data and automation while drawing from established civil rights and ethical principles to frame their response.

Few Core Ideas

- Fairness
- Explainability
- Auditability
- Accuracy

The principles above are a step toward fair, accountable, and transparent algorithms.

Fairness - Including fairness itself as a principle is a direct way to address a core concern around data-driven decision making — that the decisions are unfair. However, conversations around fairness can quickly become fraught as each participant attempts to define fairness based on their particular background and values. While it's difficult to pinpoint what determines whether an algorithm is "fair," the problem of bias in automation requires us to demand fairness based not only on the process, but also on the results.

A fair system or design is demonstrably not biased against individuals or groups of people, but fair outcomes are difficult to prove. Validating that the results of a given equation are fair requires knowing some information about the process and a representative sample of the outcomes it produces.

Some Questions to be asked by designers

- Are there particular groups which may be advantaged or disadvantaged, in the context in which you are deploying, by the algorithm / system you are building?
- What is the potential damaging effect of uncertainty / errors to different groups?

Explanation-

The obscurity of decision making is a longstanding policy issue in the U.S., which is why there are laws requiring explanations; for example, in instances of adverse credit decisions. But technology makes it harder to understand the connections between inputs and outcomes. Developers can't always explain the reasons why their technology gave a particular result, especially when they utilize sophisticated forms of algorithmic decision making that rely on machine learning or neural networks. This is problematic for any principle that proposes a right to explanation for how an automated decision was made about a person — sometimes the question just cannot be answered, and sometimes the answer would be incredibly burdensome to calculate and would require technical expertise on the part of the individual to understand.

Also problems with explanation- some algorithms are proprietary and may give out profitable details. Private companies have a huge financial incentive to resist sharing details that might compromise their products. For example, the U.S. Association of Computing Machinery asks that "institutions that use algorithmic decision-making …produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made." Far from a call to provide in-depth technical details, this principle is focused on increasing transparency around two things: the process and the results. Their statement asks that institutions "ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms." In both frameworks, an institution could fulfill an expectation of explainability with a general sense of what factors are considered in a decision and, if applicable, what qualities of the individual affected were included in the decision.

EXAMPLES-

This extends to informing people (in plain language) when the data collected or inferred about them will be used as the basis for automated decisions. For example, Fair, Isaac and Company (FICO) provides a general sense of the factors that feed into their credit scoring algorithm as well as the weights given to them. This helps individuals understand how to prove themselves within the system and allows for a cultural dialogue about what it means to be creditworthy as determined by automated systems.

Google has shared some details about how their PageRank algorithm weights various factors in determining search results. (It's worth noting that Google has additional protection because this technology is patented, rather than protected by trade secrets. Companies might consider the ability to offer increased transparency as one of the benefits of other forms of IP protection.)

Auditability

Audits are one method to provide explanations and redress without compromising the intellectual property behind the business model.

Audit Proposal - Creating a system that can be audited creates accountability and credibility, particularly if the result of an audit can be reviewed externally. The U.S. ACM proposes that creators of automated decision making have a responsibility to record their "models, algorithms, data, and decisions …so that they can be audited in cases where harm is suspected." Some have taken the call for audits further and proposed that systems be accessible by third-parties that can "probe, understand, and review the behaviour of the algorithm through disclosure of information that enables monitoring, checking, or criticism." Anticipating and complying with the needs of a large-scale, external audit can help companies detect and mitigate discriminatory impacts of their technology.

How to embed these principles with the technology

Principles established by academics, advocates, and policymakers are meant to demonstrate a philosophy that should be embedded throughout automated systems. This puts the burden on designers to understand how to integrate the goals of the principles into the technology itself.

Explained in the graphic

Resources

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

https://cdt.org/issue/privacy-data/digital-decisions/

https://drive.google.com/file/d/0B-wQVEjH9yuhanpyQjUwQS1JOTQ/view

https://www.fatml.org/resources/principles-for-accountable-algorithms