A.I. 'BIAS' DOESN'T MEAN WHAT JOURNALISTS SAY IT MEANS

Chris Stucchio & Lisa Mahapatra - August 29, 2017 -

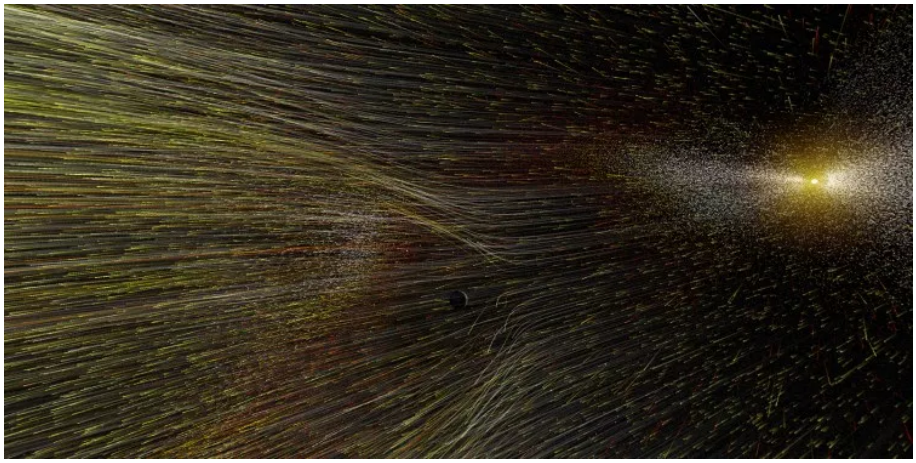


Image source

In Florida, a criminal sentencing algorithm called COMPAS looks at many pieces of data about a criminal and computes the probability that they will commit new crimes. Judges use these risk scores in criminal sentencing and parole hearings to determine whether the offender should be kept in jail or released.

In 2016, ProPublica wrote an article accusing this algorithm of being biased against blacks as well as being unreliable and frequently making errors. Like many, I read this article, shocked and horrified that such an algorithm could be used. I dug into their R-script, which is the computer program ProPublica used to analyze the data. The results of the statistical test they ran was negative — i.e. there was no evidence that the algorithm treated blacks any differently from whites. The same statistical test showed that people rated as "High Risk" by COMPAS were significantly more likely to re-offend than people rated as "Low Risk." Moreover, at any given risk level, blacks and whites had similar probabilities of re-offending.

I was very confused at this point — did I look at the wrong data analysis? (For more details on this, see here.)

COMPAS is not the only computer algorithm that has been accused by journalists of being biased. Google adwords was accused of unfairly showing ads for high paying jobs to men. Financial algorithms for predicting loan repayment are regularly accused of being racist. The media regularly decries algorithms as "biased," "sexist," and "racist."

As a person who knows how algorithms work, this is very surprising to me. Algorithms

THE LANDSCAPE OF INNOVATIVE GOVERNANCE

Eden Hardcastle - March 23, 2018

ARCHIVES

June 2018 May 2018

As a person who knows how algorithms work, this is very surprising to me. Algorithms are nothing more than a repeatable mathematical procedure, executed by a computer. Think of the step-by-step directions you were given in grade school for solving systems of linear equations or the quadratic formula. How could such a formula suddenly become biased against, say, football players?

The media is misleading people. When an ordinary person uses the term "biased," they think this means that incorrect decisions are made — that a lender systematically refuses loans to blacks who would otherwise repay them. When the media uses the term "bias", they mean something very different — a lender systematically failing to issue loans to black people regardless of whether or not they would pay them back.

Before getting into that, I will need to spend some time explaining what statistics (or "data science" and "AI" if we want to use marketing terms) is all about.

What is data science and AI?

Statistics is the mathematical field of programming computers to look at large amounts of data, and then finding hidden patterns in the data to predict future data points that I haven't seen.

Here's an example of the kind of things data scientists do. This example is fictitious (in order to preserve the confidentiality of my clients) but representative of the sort of work data scientists do.

I was asked to build a statistical model — or "AI" as marketers like to call it — to predict whether a visitor to a certain website was likely to make a purchase. If they were, we would give them a special offer. But special offers cost money, so we didn't want to make special offers to people unlikely to purchase.

Here's how I build the model: I took all the past data (about site visitors, pages viewed, etc) big spreadsheet — one column representing whether they purchased, the other columns representing attributes we think might be useful in predicting this.

Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Purchased?
विहार	Chrome	False	हिंदी	5	False
महाराशत्र	Chrome	False	मराठी	5	True
महाराशत्र	Firefox	True	हिंदी	5	True
विहार	Firefox	False	UK/English	8	False
...	...	...	...	...	...

The "feature" columns represent data used to predict the outcome; e.g., "Feature 3" might represent whether the user came from Reddit. The "Purchased?" column represents what the model is trying to predict — whether the visitor made a purchase or not. The model (or "AI") then does a bunch of math on the whole spreadsheet and looks for patterns.

An astute reader who is good at spotting patterns might notice something in this data. Whenever "Feature 1" is "महाराशत्र" the user makes a purchase, while when "Feature 1" is "बिहार" they do not. Most likely you, the reader, can spot this pattern even if you have no idea what "बिहार" means.

In real life it would be silly to assume this is true having seen the pattern only 4 times — it could easily just be a fluke. Humans are prone to drawing inferences too quickly in cases like this. But a statistical model is designed to avoid this error, and only draw inferences when the data set is large enough to support the conclusion.

But nevertheless, after looking at millions of rows, my AI has concluded that "Feature1 = बिहार" implies "less likely to purchase", and "Feature1 = महाराशत्र" implies "more likely to purchase." That's far less likely to be a fluke.

More realistically, our predictor might be a more complicated formula like:

April 2018 March 2018 Top
February 2018 January 20
December 2017
November 2017 October
September 2017 August 2
July 2017 June 2017
May 2017

purchase_probability = 12% x "Feature1=महाराशत्र" + 3% x Feature2=Chrome + 4% x Feature4 = English + etc.

(In reality, far more complicated predictors are possible. But I'm keeping things simple.)

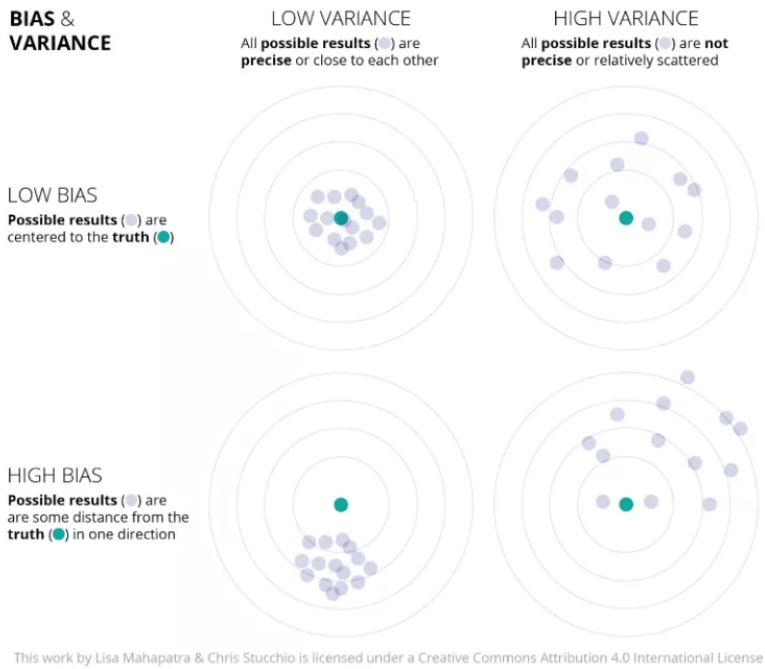
Now after I learned Hindi I realized that "बिहार" actually means "Mobile." So what our predictor discovered is that people on a mobile phone are less likely to make a purchase. So this result is hardly surprising.

What is bias?

The term bias originally refers to a "slant" of the lawn on which the game "bowls" was played. The effect of a bias is that if two teams were evenly matched, the team favored by the slanted playing field would be more likely to win.

In statistics, a "bias" is defined as a statistical predictor which makes errors that all have the same direction. A separate term — "variance" — is used to describe errors without any particular direction.

It's important to distinguish bias (making errors with a common direction) from variance which is simply inaccuracy with no particular direction. The following diagram illustrates:



The effect of bias is very simple. If the predictor we described above is biased against "Feature1 = विहार", this means we are incorrectly labelling some people as "low probability of sale" when they have a high probability of a sale. This means we're leaving money on the table by not giving them the special offer!

On the flip side, if the predictor is biased in favor of "Feature1 = विहार", then we are wastefully issuing special offers that will never be redeemed. That also costs us money.

How do we identify bias?

To identify bias in a predictor, one straightforward method is to build a new predictor which removes what we believe is bias. Then we check whether the new predictor gives us a more systematically accurate prediction.

Let us go back to the example above, where "Feature1 = महाराशत्र" makes us believe a purchase is 12 percent more likely. In this case, we could compare that predictor to an alternative (hypothetically) unbiased predictor which ignores Feature 1.

If we are right about this bias, then our new model will make more money because we'll get more sales. And if we are wrong it will cost us money. The key point is that bias is purely mathematical.

In this case, I ran this analysis and determined that my algorithm was not biased.

The fallout

Sometime after I deployed this algorithm, my client got some angry calls. People in Bihar were very annoyed that the website kept giving special offers to people in Maharashtra but not to them. An angry newspaper columnist criticized the company: "your algorithm is biased, it doesn't recognize that Biharis also purchase things!" Local social justice activists criticized the company on the same grounds, called me a shivsainik (a supporter of the Shiv Sena political party) and accusing me of racism against non-Marathis.

This was a bit silly; for those unfamiliar, Shiv Sena is a racist political party in Maharashtra state (where I lived) which wants to prevent non-Marathis (like me) from immigrating and finding jobs. Needless to say I have no affiliation with them apart from occasionally working out at a free gym that they operate.

What happened? It turns out that I don't speak very good Hindi. विहार actually refers to the Indian state of Bihar and महाराशत्र is the Indian state of Maharashtra. What I thought meant "mobile phone browser" and "desktop browser" actually meant "visitor is from Bihar" and "visitor is from Maharashtra." But the meaning of these terms doesn't change the math at all. The AI algorithm is making correct predictions either way.

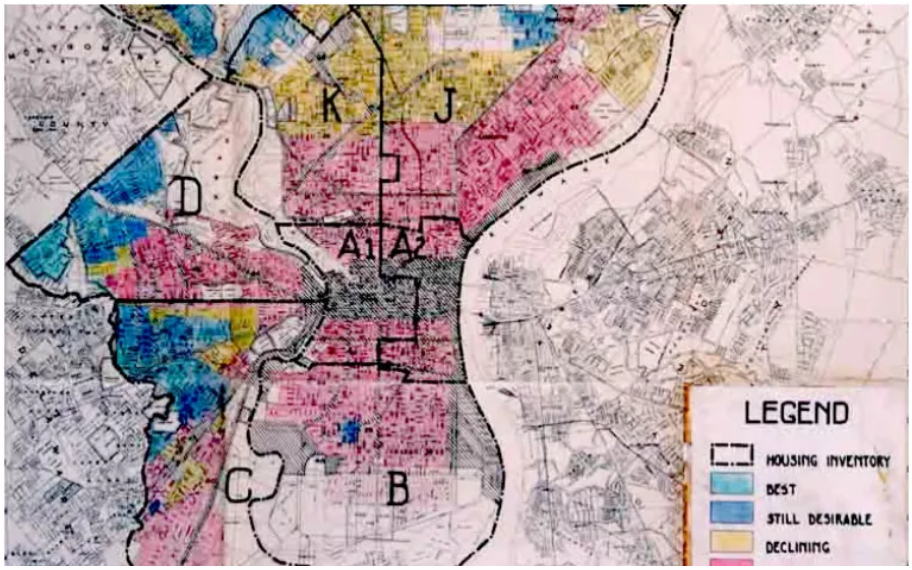
It turns out that without even meaning to, I had discovered an uncomfortable truth. Maharashtrians are likely to make a purchase, but Biharis are not.

What do journalists think "bias" refers to?

What I've described above is bias according to a statistician — the difference between an algorithm's average prediction and reality. But journalists define "bias" in a very different way. I'll provide some examples.

Predicting creditworthiness

The practice of redlining began with the National Housing Act of 1934, which established the Federal Housing Administration (FHA) and provided subsidies for mortgages. The FHA drew maps with "red-lined" areas — which it labelled "Hazardous" — which were subject to lower levels of subsidies. These areas were often predominantly black, which effectively meant that the FHA subsidies were being given primarily to non-black people.



As a response to this practice, financial regulators now impose a variety of complex and non-deterministic requirements on lenders relating to algorithms making lending decisions.

An issue that is arising in the modern world is that statistical algorithms often reproduce red lines. As Delip Rao discusses, even if a model does not know the race of loan applicants, if it has enough input data it might be able to infer it. For example, data like "income < $15k, location = Oakland" is highly correlated with being black.

Some researchers at Google studied this issue. They discovered that Blacks and Asians with the same FICO score have significantly different debt repayment rates — an Asian person with a FICO of 600 has an 80 percent chance of repaying a loan, while a Black person with the same FICO has only a 60 percent chance. (The authors refused my request to provide the original data, so my numbers were obtained from the graph using a ruler and eyeballing graph ticks.) So a profit maximizing lender making use of this data would charge blacks a much higher interest rate than Asians.

In Delip Rao's presentaton, he points out that if race is predictive of defaults, and data contains enough information to make inferences about race, then a sufficiently accurate model might deduce the race of a loan applicant and reject them for a loan.

It's very important to note that although journalists describe this as a "new kind of redlining", it is not actually an incorrect decision. The same Google researchers above created a very nice interactive page illustrating the results of their paper. The page illustrates that the most accurate and profitable algorithm directly discriminates by race, and constructs other "fair" algorithms which also discriminate by race in ways the authors consider more "fair." The result is more bad loans are issued and profits are reduced.

Understanding Language

Here is an example: Language necessarily contains human biases, and so will machines trained on language corpora. In this article (and the corresponding academic paper), an algorithm is taught human language. The algorithm then draws a semantic association between certain professions and the female gender. The result is here:

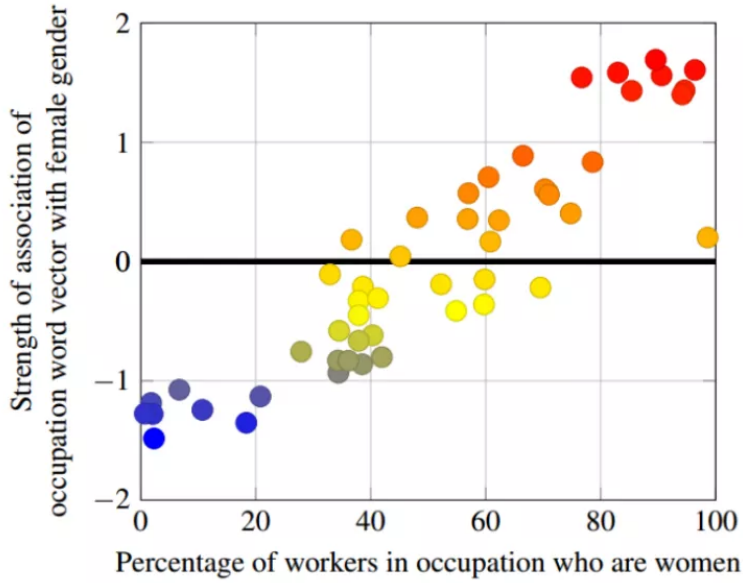


Figure 1. Occupation-gender association
Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

(The algorithm can also accurately infer the gender of specific people with androgynous

names.)

So to recap, an AI has read a bunch of text on the internet and accurately learned a true fact about reality: there are lots of female nurses, and very few female physicists. This is termed "bias."

Predicting Criminal Behavior

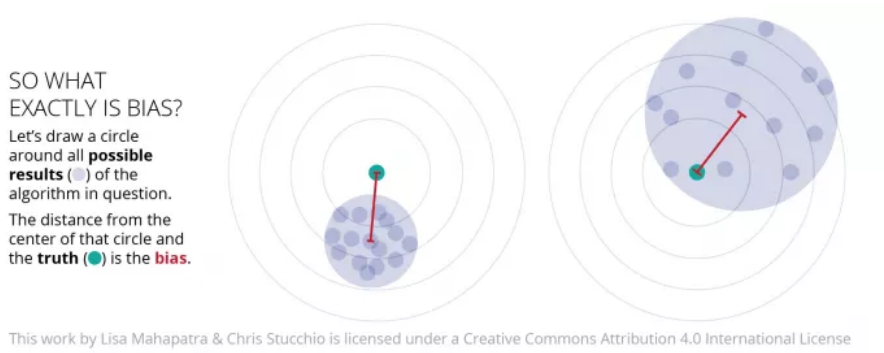
As discussed in the beginning, ProPublica claimed that the COMPAS algorithm is biased against blacks. ProPublica labelled the algorithm as biased based primarily on the fact that it (correctly) labelled blacks as more likely than whites to re-offend (without using race as part of the predictor), and that blacks and whites have different false positive rates. This is actually just a necessary mathematical consequence of having an unbiased algorithm — no decision process, whether implemented by an AI or a sufficiently diverse group of humans, could possibly avoid this tradeoff.

In the conception of these authors, "bias" refers to an algorithm providing correct predictions that simply fail to reflect the reality the authors wish existed.

This example is also important because of its real world consequences. After the article was published, a team of statisticians at Stanford decided to study the cost of fairness. It is possible to take the COMPAS algorithm and manipulate it to be fair. But in the process of doing this, accuracy is reduced. The Stanford team shows that if this were done, the (mostly black) high risk convicts that the manipulated algorithm would release would then commit 9 percent more violent crimes. Furthermore, 17 percent of the people in jail would be (mostly white) individuals at a very low risk of re-offending.

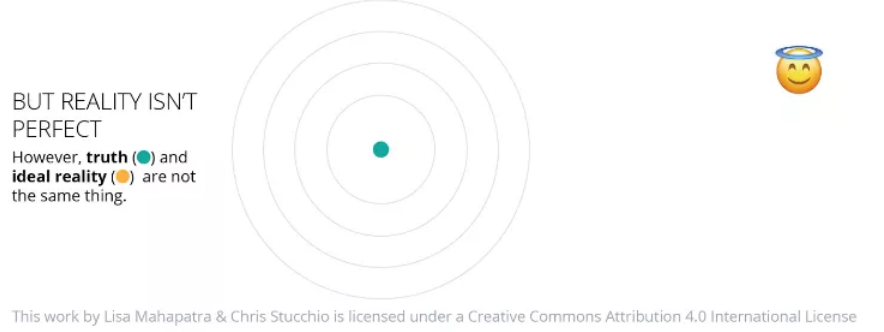
Bias according to journalists: the difference between an AI's output and wishful thinking

To a statistician, bias is defined as the difference between an AI's "typical" output and reality. Bias has a magnitude and a direction, and is a systematic tendency to get the same kind of wrong answers.



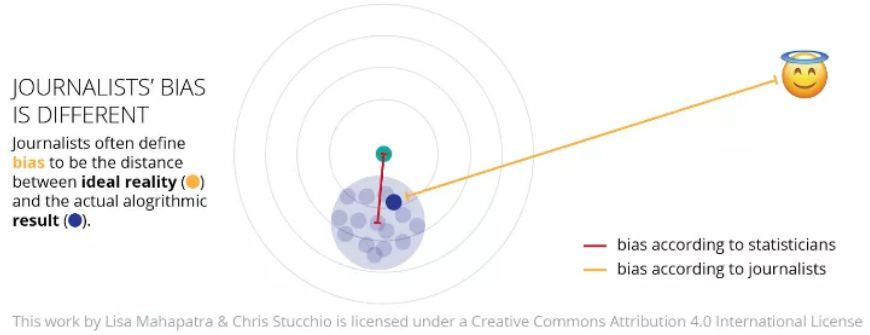
Very importantly, "wrong" is defined as being relative to reality.

One problem with reality is that it often fails to live up to our desires and expectations. There is often a very large gap between the two.



When a journalist discusses bias, they typically do not mean it in the same manner

When a journalist discusses bias, they typically do not mean it in the same manner that statisticians do. As described in the examples above, a journalist typically uses the term "bias" when an algorithm's output fails to live up to the journalist's ideal reality.



Perhaps for a journalist whose primary interest is to spark conversation, conflating these two types of bias is fine.

However, there is a significant cost to forcing algorithm outputs to reflect wishful thinking. If we are issuing loans, we will issue more loans to people who do not pay them back. If we are making parole/sentencing decisions for convicted criminals, we will release more dangerous criminals who go on to commit new crimes.

This may not be a concern for the journalist, but it should be a concern for the rest of us.

Chris Stucchio is a data scientist (formerly known as "statistician") who gambles a bit and it mostly works out. He no longer identifies as a physicist.

Lisa Mahapatra is a maker of data visualizations and data illustrations. She is a purveyor of her craft at Fintech startup dv01. She no longer identifies as a journalist.

SAVONAROLA

By Chris Morgan - July 26, 2017 -

EDGY

By Edward Waverley - May 26, 2017 -

SOCIAL MEDIA