We Are AI: Taking control of technology
We Are AI #4: All about that Bias

Cover-alt

A majestic golden mirror stands in an empty room – We see over the shoulder of an embodied AI/robot gazing into the mirror. The reflection in the mirror shows a myriad of human faces: in the front are three caucasian men, smiling confidently and beaming in the reflection. Behind them stand Asian and African-american men and women with gags over their mouths, fading into the background. The mirror exudes colorful light, drawn as abstract line art, leading from the reflection towards the AI/robot. The reflection of the AI as people in the mirror– both oppressed and empowered, is symbolic of the idea that machine bias is merely a reflection of the human biases that technology encodes.

Cite as:
Julia Stoyanovich and Falaah Arif Khan. “All about that Bias”.
We are AI Comics, Vol 4 (2021)
https://dataresponsibly.github.io/we-are-ai/comics/vol4_en.pdf

Page 1

Let's talk about what we mean by 'bias' in AI, and how it arises.

Picture a robotic head – with Googly eyes, two gray antennas on its head and keyboard buttons for teeth. In the eyes, in swirly writing is written the word "bias".

We say that an AI is biased if its use can lead to systematic and unfair discrimination against some individuals or groups in favor of others.

Bias can stem from harmful patterns picked up from the data itself, or from how the algorithm is designed, or from the objectives that we specified for it, or from how we use it.

In their seminal 1996 paper, Batya Friedman and Helen Nissenbaum identified three types of bias that can arise in computer systems, A caricatured Batya Friedman and Helen Nissenbaum stand smiling at the camera/reader.

represented here as a three-headed dragon:
pre-existing
technical
emergent
A jet-black dragon, with three heads stretches its talons towards the reader and opens its mouth in a growl – the leftmost head is facing towards the left, and has "preexisting" written across its neck in blue. The middle head is turned upwards and is roaring defiantly, and has the word "technical" written across its neck in purple. The rightmost head is facing the reader with its mouth open in a growl to reveal a slimy green tongue, and has the word "emergent" written across its neck in pink.

[1] Batya Friedman and Helen Nissenbaum. (1996). Bias in computer systems.

Page 2

Recall the baking metaphor we used to understand data-driven algorithms in Volume 1.

Let's now use the same metaphor to understand bias!

Welcome back our protagonist Mo. Mo is wearing yellow oven mitts and is taking out a loaf of bread from the oven. The loaf is a perfect golden brown on the outside and has detailed crust-work in the form of smiling robot heads.

Pre-existing (in the data)

Pre-existing bias exists independent of the algorithm and has its origins in society.

These would be the flavor notes that will seep into your bread if you don't prioritize the purity/freshness of your ingredients,

or if you decide to use premixed off-the-shelf batter.

Mo is shopping at a grocery store – she holds a basket full of ingredients and is in the 'ready to eat' aisle, picking out a box of bread-mix.

These biases exist in society and come 'pre-baked' into the algorithm

from the underlying discriminatory system that the data was collected from -

such as the gender and racial stereotypes that language models pick up when trained on data from social media.

A robot head is scowling at the viewer. Its eyebrows are furrowed with anger and it's eyes are red and raging. In its pupils we see the face of a man who is shouting. Behind the robot stands the very man, with his eyebrows crossed in anger, his hair spiked up, and his mouth mid-scream. Out of his mouth are emerging little bubbles that are feeding into the back of the robot's head. The bubbles have the logos/symbols of popular social-media platforms such as a vectorized camera and a twittering bird.

Page 3

Technical (in the technical system)

Technical bias is introduced by the system itself - because of the way it is designed or operates.

These would be the imperfections that will seep into your bread if you use the wrong equipment - such as uneven cooking of your cupcakes if your oven temperature is miscalibrated, or spillage of batter if your baking equipment is of the wrong size.

Two snapshots from Mo's baking fiascos – on the left is the picture of an overflowing baking tin, where Mo has clearly overfilled the container with batter, which has caused it to spillover during the baking process. On the right is a snapshot of some curious cupcakes – all of the cupcakes have a tear in their tops and have risen asymmetrically – they are flattened and burned on one side, but curiously fluffy and risen on the other. The batter in the middle of the tear is uncooked!

Back to computer systems:
a prominent example is social media platforms designed to optimize for engagement (instead of safety or authenticity) - that end up promoting polarizing articles and fake news.

A man is sitting, lost in daydream, with his arms folded, elbows resting on a table, his hands holding his face. Above his head is a cloud of squiggly abstract art that is emanating from the left hand of an AI-genie. The AI-genie conjures up a network of heads, spewing the same squiggly clouds that hold the man's attention – each of the heads is identical, and is screaming vitriol at the other with closed eyes, furrowed eyebrows and angry snarls.

Page 4

Emergent (due to decisions)

Emergent bias arises over time, because the decisions made with the help of the system change the world,

which in turn impacts the operation of the system going forward.

Think about behavioral changes that will emerge as a result of your baking -

What if you become such a maestro at baking that you inadvertently make bread a steady part of your diet!

Mo is holding a plate of freshly baked buns in one hand, and is stuffing her mouth full of the bakes with the other.

Or make it so often, that you turn everyone around you off the thought of ever eating another slice!

A young girl scowls at an extended plate of sourdough bread, and pushes it away with her hand.

Or think about how your idea of ‘what bread should taste like’ is shaped by the popularity of products like ‘Wonder Bread’.

A young boy – with rosy cheeks and dark curly hair, smiles at the reader, while holding up a slice of white break to his mouth. A loaf is sitting in front of him, perfectly positioned to show the label “Wonderbread” to the reader.

In the same vein, think about how your exposure to news - and information more broadly -

is shaped by algorithms that curate social feeds with popular and ‘trending’ posts.

Mo is staring up at a sky full of data, hashtags, forecasts, predictions and analytics.

Page 5

To make our discussion concrete, let’s look at real-world examples of algorithmic bias.

Let’s take ‘Hiring’ as a representative domain in which algorithms are increasingly being used to make critical decisions more ‘efficiently’.

An algorithmic hiring tool/ AI is conducting an interview – three applicants sit across a robot on a table, nervously polite and eager to impress

One of the earliest indications that there is cause for concern came in 2015, with the results of the AdFisher study out of Carnegie Mellon University [2].

Researchers ran an experiment, in which they created two sets of synthetic profiles of Web users who were the same in every respect — in terms of their demographics, stated interests, and browsing patterns — with a single exception: their stated gender, Male or Female.
Picture a splitscreen image of a prospective candidate – on the left panel is a businessman in a dapper suit, holding a briefcase, in the right panel is a businesswoman in a power suit, holding an identical briefcase. Both people in the panels are identical in their hair color, skin tone, the hue of their suits, the color and size/shape of their briefcases.

Researchers showed that Google displayed ads for a career coaching service for high-paying executive jobs far more frequently to the male group than to the female group.

This brings back memories of the time when it was legal to advertise jobs by gender in newspapers. This practice was outlawed in the US in 1964, but it persists in the online ad environment.

It was later shown that part of the reason this was happening is the mechanics of the advertisement targeting system itself, as an artifact of the bidding process. This is technical bias in action!

[2] Women less likely to be shown ads for high-paid jobs on Google, study shows. Guardian (2015)

Page 6

Let us move forward in time, and also advance to the next stage of the hiring funnel: resume screening.

An embodied AI/robot peers into a crystal ball, trying to read the future of the pool of applicants – their heads swirl around in the crystal ball, and the AI peers deeply into it, with its hands stretched around it.

In late 2018 it was reported that Amazon’s AI recruiting tool, developed with the stated goal of increasing workforce diversity, in fact did the opposite thing: the system taught itself that male candidates were preferable to female candidates.

It penalized resumes that included the word “women’s,” as in “women’s chess club captain.”
And it downgraded graduates of two all-women’s colleges.

The results aligned with, and reinforced, a stark gender imbalance in the workforce.
This is emergent bias in action -
A hiring manager to whom an AI tool repeatedly suggest the
same kind of job applicant as a good fit,
will over time come to believe that this is what a promising employee looks like.
Two caucasian men pose at the reader – the one on the left is doing a ‘thumbs down’ sign, the one on the right is doing a ‘thumbs-up’ sign. The one on the left has the symbol of a woman – a circle with a small cross underneath – on his thumb, whereas the one on the right has the symbol for a man – a circle with an arrow pointing northeast – on his thumb. Behind them are mirror reflections of the two men, but instead of their silhouettes, embodied AIs/robots are pulling the same thumbs up and thumbs down poses. There are 3 sets of such reflections, and at the very top is a reflection of the two men doing the same poses, symbolizing that pre-existing bias is encoded and then exacerbated by technology, and then presented as “neutral” human intuition/judgment to the decision-maker.

We are also seeing pre-existing bias in this example: the AI tool was trained on historical data about past employees, who were predominantly male.

[3] Amazon scraps secret AI recruiting tool that showed bias against women. Reuters (2018)

Page 7

Here’s another example, later yet in the hiring process, perhaps during a post-interview background check by a potential employer -
Latanya Sweeney, a computer science professor on the faculty at Harvard,

A caricatured Latanya Sweeney is smiling at the reader knowingly, and pointing upwards at the results of her study.

showed that Googling for African-American sounding names is more likely to trigger ads suggestive of a criminal record than googling for White-sounding names, even controlling for whether an individual in fact has a criminal record!

There are two sets of women's faces – to the top-left is a set of Caucasian women's faces – with blonde hair and fair skin. Below them is a magnifying glass – symbolic of "search", and the name "Kristen" is written across it. To the bottom-right is a set of African-American women's faces – with dark, curly hair and a dark complexion. Below them is the "search" symbol, with the name "Latanya" written across it.

This is pre-existing bias at play - manifesting long-standing racial prejudices of society.

[4] Racism is Poisoning Online Ad Delivery, Says Harvard Professor. MIT Technology Review (2013)

Page 8

The cases presented here have one thing in common: they show that AI can reinforce and exacerbate unlawful discrimination against minority and historically disadvantaged groups.
Often this is called out as "bias in AI".

The same robotic head – with Googly eyes, two gray antennas on its head and keyboard buttons for teeth has re-appeared in our discussion of bias now. In its eyes, in swirly writing is written the word "bias".

So, why are sophisticated systems that aim to make hiring more efficient failing at this, and arguably making things worse?

Of course, the issues of bias in employment are not new. They exhibited themselves in the analog era as well.

For example, in their well-known 2004 study, Marianne Bertrand and Sendhil Mullainathan sent fictitious resumes to help-wanted ads in Boston and Chicago newspapers. [5]

Are Emily and Greg more employable than Lakshika and Jamaal?

To manipulate perceived race, they randomly assigned African-American- or White-sounding names to resumes.

White names receive 50 percent more callbacks for interviews.

Caricatured Marianne Bertrand and Sendhil Mullainathan look at the reader, while pointing towards a large box above them. In the box are the words “Are Emily and Greg more employable than Lakisha and Jamal?”. To the left of the text are the faces of a Caucasian male and female. To the right are the faces of an African-American man and woman.

This case shows that bias can be due to human decisions.

[5] Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. Marianne Bertrand&SendhilMullainathan (2003)

Page 9

Let’s revisit pre-existing bias that often exhibits itself in the data.

Data is an image of the world, its mirror reflection.

When we think about bias in the data, we interrogate this reflection.

One interpretation of “bias in the data” is that the reflection is distorted - we may systematically over-represent or under-represent particular parts of the world in the data, or otherwise distort the readings.

A woman sits at her computer and we get to see over her shoulder: In front of her, on her laptop a spreadsheet is open, and from the laptop arises fragments of data. The data is coalescing into a round shape that is sitting flat on top of a pink mirror. The round shape is a pixelated reflection, and above it we see the original image of the world – a circular globe drawn with abstract line art.

Recall the failure of Amazon’s recruiting AI to improve workforce diversity.

This tool was trained using historical data: resumes of people who were hired in the past.
That training was subject to pre-existing bias.

In that data, there was an under-representation of women in the workforce, and in technical roles.

A more subtle point is about distortions.

When we consider features, like an individual's score on a standardized test, do we take these at face value?

Or do we account for differences in access to educational opportunity, like going to a better school, or having access to paid tutoring?

Page 10

Another interpretation of "bias in the data" is that even if we were able to reflect the world perfectly in the data, it would still be a reflection of the world such as it is,
and not necessarily of how it could and or should be.

It is important to keep in mind that a reflection cannot know whether it is distorted.
Data alone cannot tell us whether it is a distorted reflection of a perfect world, a perfect reflection of a distorted world, or if these distortions compound.

A woman stares at what appears to be a portrait / art installation of some kind. The portrait is a 3-D semi-sphere of green and blue pixels, mounted on a pink grid. Behind it, and beyond the view of the woman is a globe drawn in abstract line art. Half of the globe – the half towards the woman is drawn in blue and green, whereas the other half is drawn filled with black over green and blue lines. In the middle of the globe is a massive question mark in red.

The second point is that it is not up to data or algorithms, but rather up to people — individuals, groups, and society at large — to come to consensus about whether the world is how it should be, or if it needs to be improved.

And, if so, how we should go about improving it.

We see the silhouette of a group of people looking at what appears to be an art installation of some sort – there are 4 different representations of a globe. From left to right: there's a realistic 3-D rendering of the different terrains and geographical elements of the earth, then an abstract drawing using only concentric polygons in bright purples and pinks, then a sketch composed of abstract line art in blue and green, and finally an even more abstract polygon art composition of different colorful shapes.

Page 11

The final point here is that changing the reflection may not change the world.

If the reflection itself is used to make important decisions - for example, whom to hire or what salary to offer to an individual being hired, then compensating for the distortions is worthwhile.

But the mirror metaphor only takes us so far.

We have to work much harder — usually going far beyond technological solutions — to make lasting change in the world, not merely brush up the reflection.

Picture the 3-D semi-sphere of green and blue pixels, mounted on a pink grid, once again. Behind it, to the far right is a globe drawn in abstract line art. Half of the globe – the half towards the semi-sphere is drawn in blue and green line art filled with black,, whereas the other half is drawn filled with white. A woman is holding cleaning liquid in one hand, and a dishcloth in the other, and is cleaning the semi-sphere (reflection) of the abstract globe (world). A man stands beside her, with a clipboard and pen, looking at the semi-sphere and taking notes.

Circling back now to the three-headed Bias Dragon.

When speaking about tackling Bias in AI, we tend to frame the problem as finding a way to slay the bias-dragon.

The three-headed bias dragon has made a re-appearance. This time, it's breathing fire through two of its three heads. A valiant knight is battling the dragon, and is piercing the middle head with a sword. However, the other two heads are dousing the knight in flames.

But through our discussion of the irrevocable link between human bias and machine bias,
we find ourselves questioning the very nature of this tale -

At the end of the day, maybe the question isn't -
how to slay the dragon and rescue the princess?
A princess looks down from the window of her castle, at a gallant knight standing atop the head of the dragon they have just slain.

The question we really should be asking ourselves is -
What do we do about a society that locks up princesses in castles, in the first place?