WE ARE AI
#4

All about that BIAS

© Julia Stoyanovich and Falaah Arif Khan (2021)

TERMS OF USE

All the panels in this comic book are licensed CC BY-NC-ND 4.0. Please refer to the license page for details on how you can use this artwork.

TL;DR: Feel free to use panels/groups of panels in your presentations/articles, as long as you

1. Provide the proper citation
2. Do not make modifications to the individual panels themselves

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

Contact:

Please direct any queries about using elements from this comic to themachinelearnist@gmail.com and cc stoyanovich@nyu.edu

Licensed CC BY-NC-ND 4.0

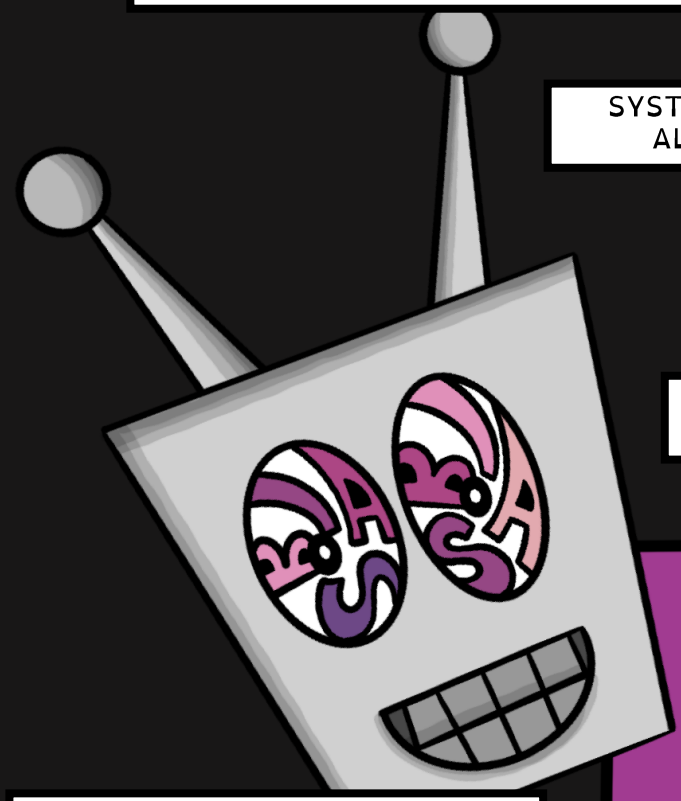
© Julia Stoyanovich and Falaah Arif Khan (2021)

LET'S TALK ABOUT WHAT WE MEAN BY 'BIAS' IN AI, AND HOW IT ARISES.

SYSTEMATIC DISCRIMINATION BY AN ALGORITHM IS TERMED 'BIAS'.

THESE COULD BE HARMFUL PATTERNS PICKED UP FROM THE DATA ITSELF,

OR FROM HOW THE ALGORITHM IS DESIGNED,

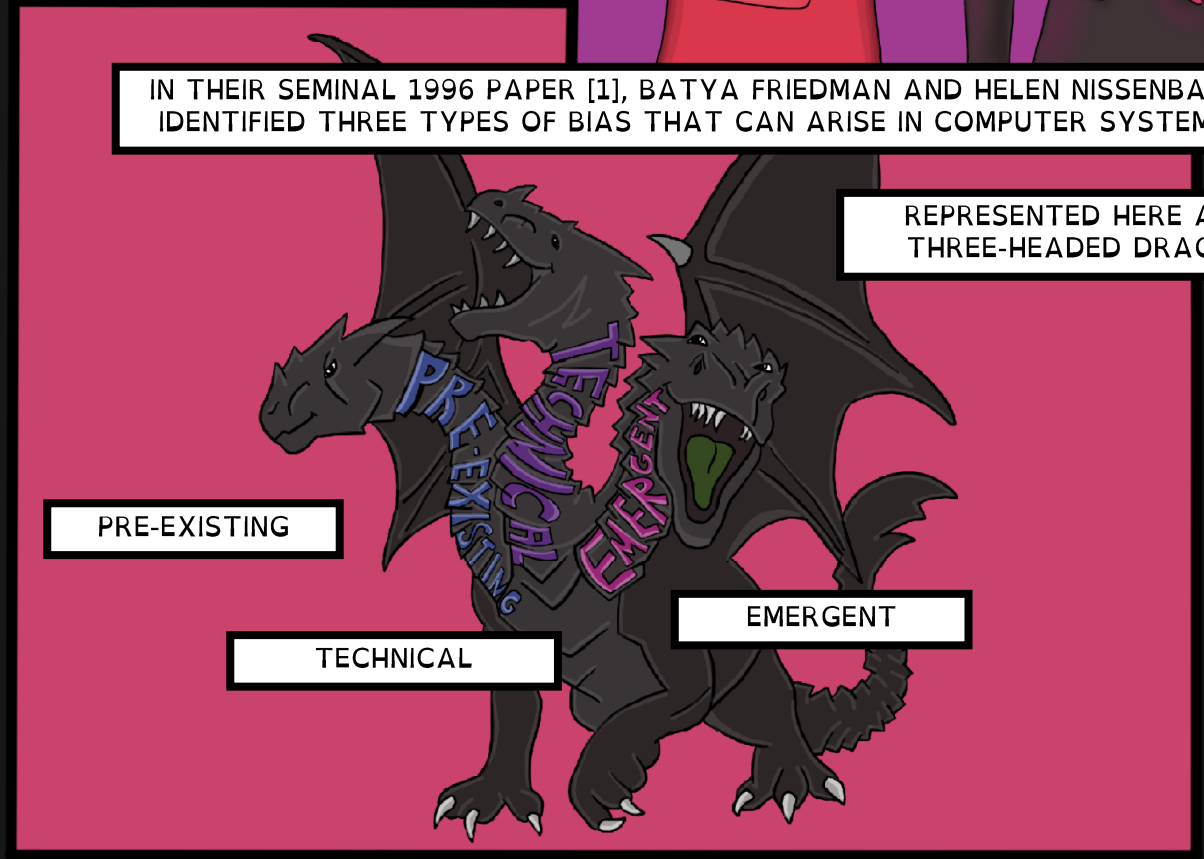


OR FROM THE OBJECTIVES THAT WE SPECIFIED FOR IT,

OR FROM HOW WE USE IT.

IN THEIR SEMINAL 1996 PAPER [1], BATYA FRIEDMAN AND HELEN NISSENBAUM IDENTIFIED THREE TYPES OF BIAS THAT CAN ARISE IN COMPUTER SYSTEMS,

REPRESENTED HERE AS A THREE-HEADED DRAGON:



PRE-EXISTING

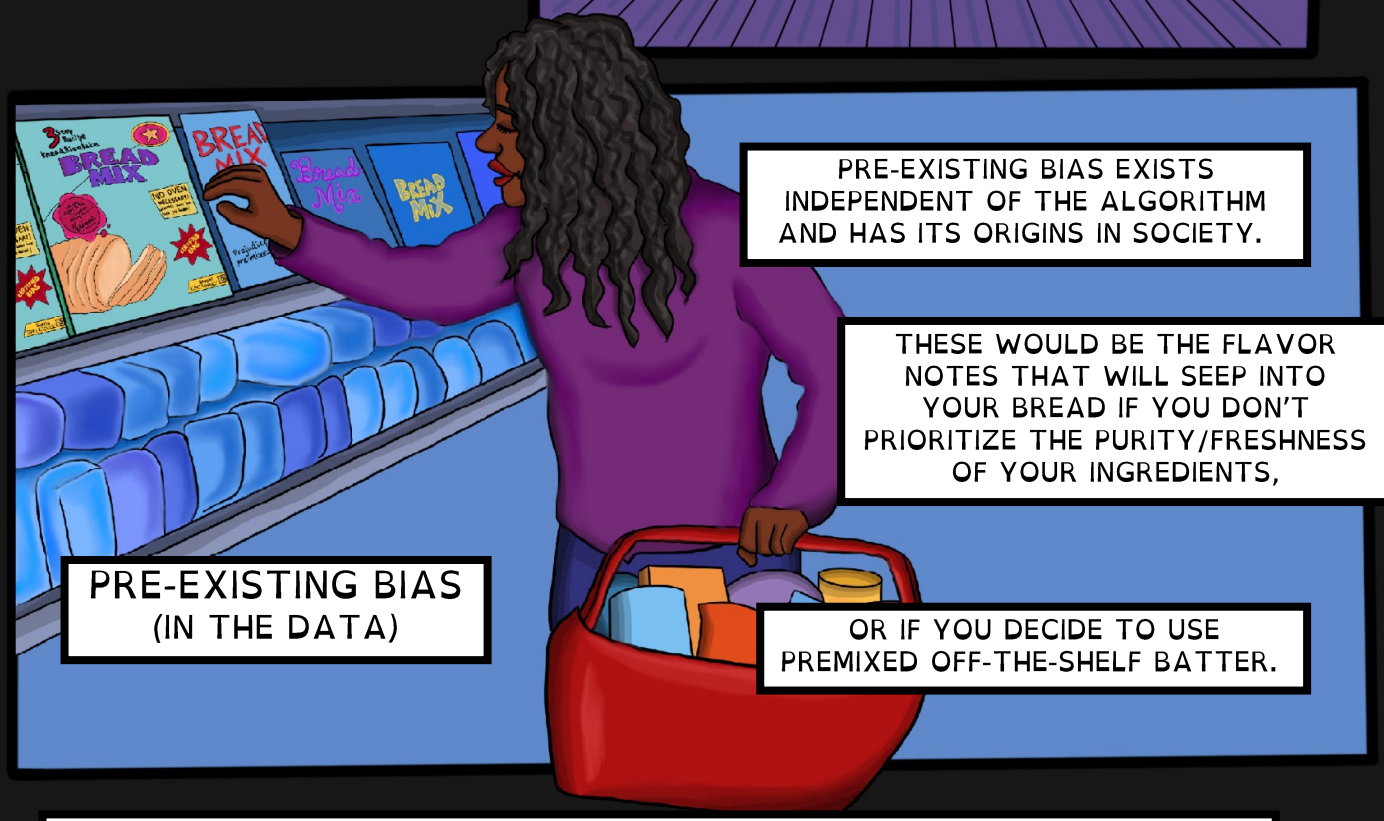
TECHNICAL

EMERGENT

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

RECALL THE BAKING METAPHOR WE USED TO UNDERSTAND DATA-DRIVEN ALGORITHMS IN VOLUME 1.

LET'S NOW UNDERSTAND BIAS THROUGH THAT SAME LENS!



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,

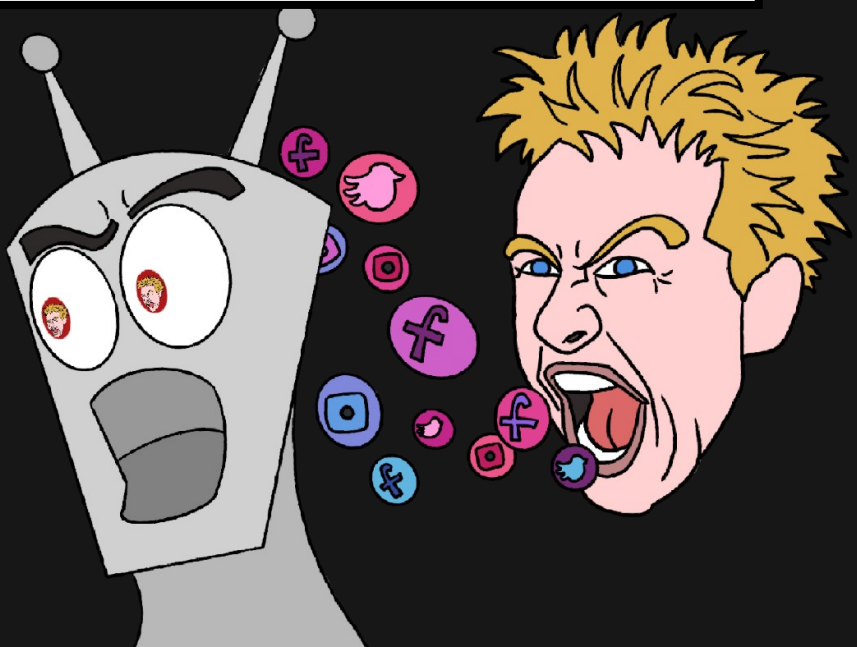
PRE-EXISTING BIAS
(IN THE DATA)

OR IF YOU DECIDE TO USE PREMIXED OFF-THE-SHELF BATTER.

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.



TECHNICAL BIAS

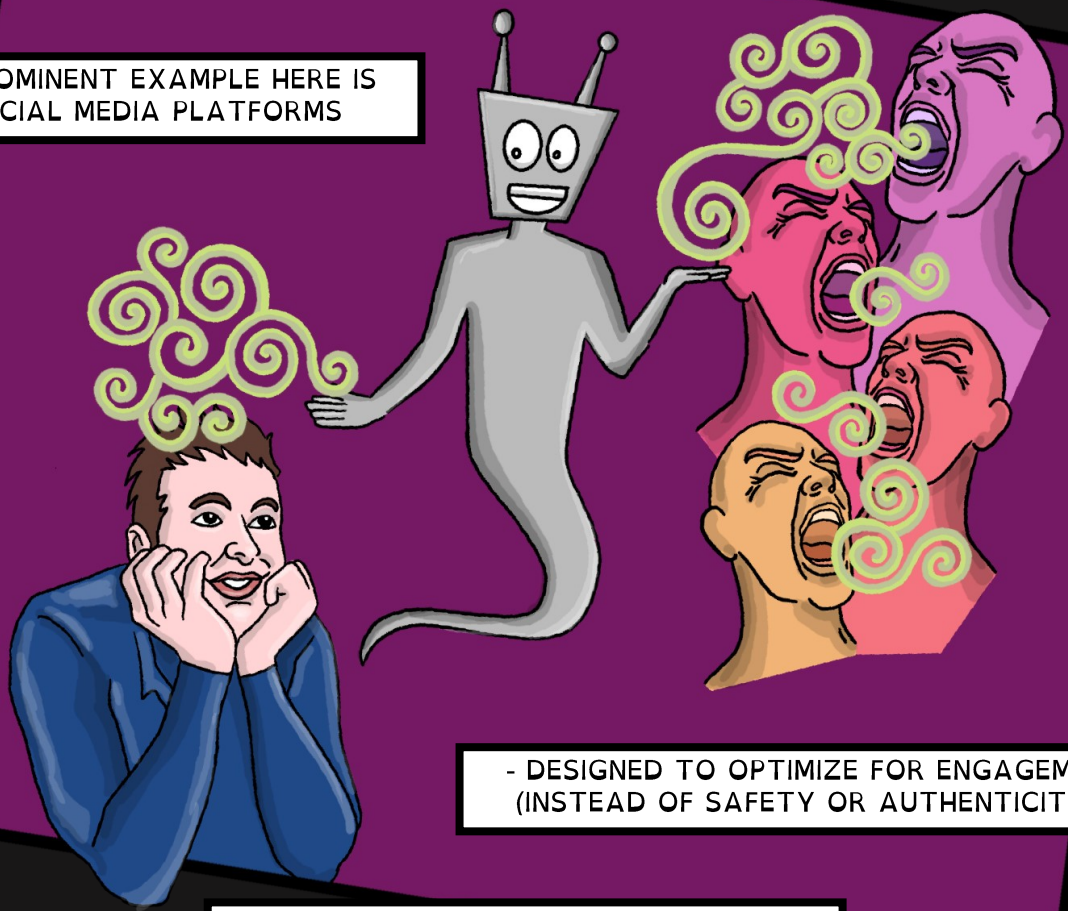
TECHNICAL BIAS IS INTRODUCED BY THE SYSTEM ITSELF - BECAUSE OF THE WAY IT IS DESIGNED OR OPERATED.

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 THE WRONG SIZE.

A PROMINENT EXAMPLE HERE IS SOCIAL MEDIA PLATFORMS



- DESIGNED TO OPTIMIZE FOR ENGAGEMENT (INSTEAD OF SAFETY OR AUTHENTICITY) -

THAT END UP PROMOTING POLARIZING ARTICLES AND FAKE NEWS.

EMERGENT BIAS
(DUE TO DECISIONS)

EMERGENT BIAS ARISES OVER TIME, BECAUSE THE DECISIONS MADE WITH THE HELP OF THE ALGORITHM 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!

OR MAKE IT SO OFTEN, THAT YOU TURN EVERYONE AROUND YOU OFF THE THOUGHT OF EVER EATING ANOTHER SLICE!

OR THINK ABOUT HOW YOUR IDEA OF 'WHAT BREAD SHOULD TASTE LIKE' IS SHAPED BY THE POPULARITY OF PRODUCTS LIKE 'WONDER BREAD'.

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.

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'.

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.

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.

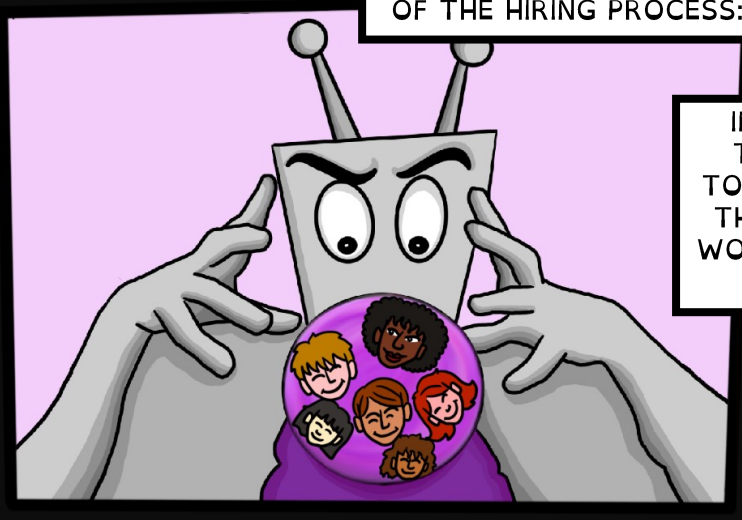
THIS BRINGS BACK MEMORIES OF THE TIME WHEN IT WAS LEGAL TO ADVERTISE JOBS BY GENDER IN NEWSPAPERS. THIS PRACTICE WAS OUTLAWED IN THE US 1964, BUT IT PERSISTS IN THE ONLINE AD ENVIRONMENT.

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)

LET US MOVE FORWARD TO THE NEXT STAGE OF THE HIRING PROCESS: **RESUME SCREENING.**



IN LATE 2018 IT WAS REPORTED THAT AMAZON'S AI RECRUITING TOOL, THAT WAS DEVELOPED WITH THE STATED GOAL OF INCREASING WORKFORCE DIVERSITY, IN FACT DID THE OPPOSITE THING: [3]

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 OVERTIME COME TO BELIEVE THAT THIS IS WHAT A PROMISING EMPLOYEE LOOKS LIKE.

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

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,

SHOWED THAT GOOGLING FOR AFRICAN-AMERICAN SOUNDING NAMES IS MORE LIKELY TO TRIGGER ADS SUGGESTIVE OF A CRIMINAL RECORD THAN WITH WHITE-SOUNDING NAMES,

Kristen

EVEN CONTROLLING FOR WHETHER AN INDIVIDUAL IN FACT HAS A CRIMINAL RECORD! [4]

Latanya

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)

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".

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 LAKISHA and JAMAL?

TO MANIPULATE PERCEIVED RACE, THEY RANDOMLY ASSIGNED AFRICAN-AMERICAN- OR WHITE-SOUNDING NAMES TO RESUMES.

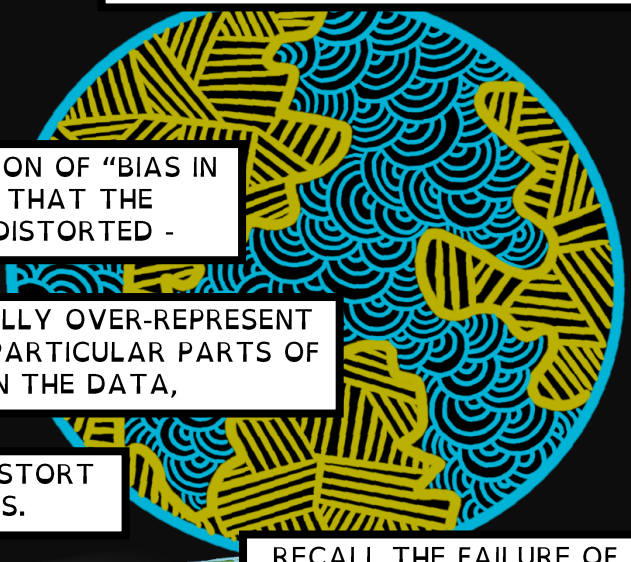
WHITE NAMES RECEIVE 50 PERCENT MORE CALLBACKS FOR INTERVIEWS.

THIS CASE SHOWS THAT BIAS IS DUE TO HUMAN DECISIONS.

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

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.

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?

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 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.

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.

A 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.

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.

PRE-EXISTING

TECHNICAL

EMERGENT

BUT THROUGH OUR DISCUSSION OF THE 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?

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?

FIN.