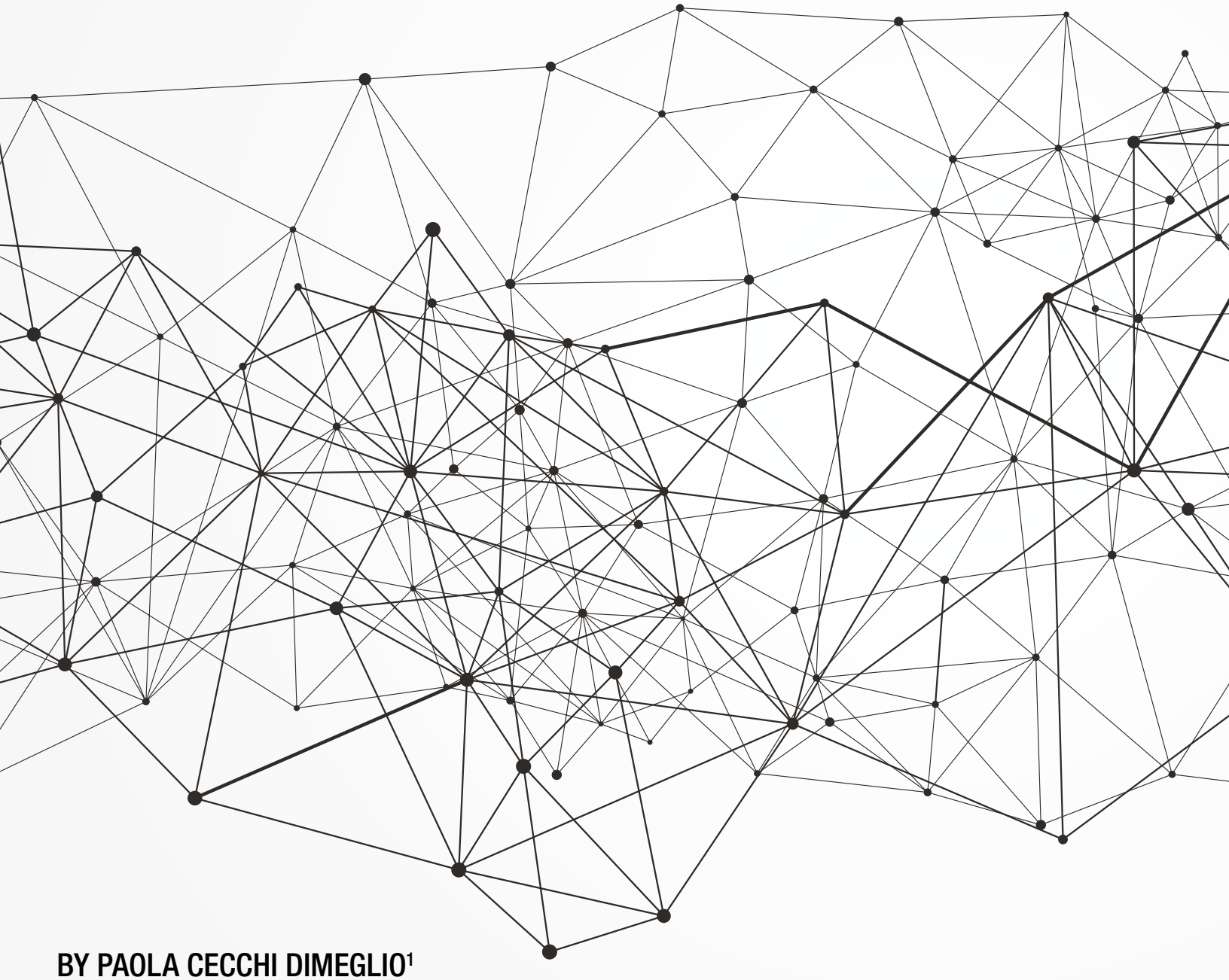
CAN WE GET THE BIAS OUT OF OUR AI?



BY PAOLA CECCHI DIMEGLIO[1]

1 Paola Cecchi-Dimeglio, a lawyer, and behavioral and data scientist, is chair of the Executive Leadership Research Initiative for Women and Minority Attorneys at Harvard Law School and Harvard Kennedy School. She is also CEO of the People Culture Drive Consulting Group and author of Diversity Dividend, forthcoming from MIT Press. Email: pcecchidimeglio@law.harvard.edu

CPI ANTITRUST CHRONICLE JUNE 2023

Visit www.competitionpolicyinternational.com for access to these articles and more!

CAN WE GET THE BIAS OUT OF OUR AI?

By Paola Cecchi Dimeglio

This article sheds light on how algorithms, originally intended to promote fairness and automation, can inadvertently perpetuate discrimination. By examining various domains such as employment, housing, banking, and education, one can uncover the far-reaching effects of bias, influencing outcomes and potentially reinforcing societal prejudices. Recognizing the urgency of the matter, the article underscores the significance of early detection and effective intervention to address algorithm bias. It highlights valuable strategies like diverse team involvement, inclusive dataset testing, and robust monitoring and review processes to identify and rectify biases. Transparency and user feedback play vital roles in mitigating bias and fostering a sense of fairness. With a collective responsibility, individuals and organizations are called upon to confront algorithm bias head-on. The aim is to forge AI systems that transcend default biases, aligning with the fundamental principles of equity and inclusivity. By embracing best practices, we can strive for a future where AI algorithms stand as unbiased pillars, actively contributing to a society that is truly equitable.

I. INTRODUCTION

Algorithm bias stacks the system against us, but we may never find out. We dread sentient, adversarial artificial intelligence ("AI"), but worse possibilities are already here. The algorithms that facilitate machine learning and drive our AI are discriminating. Meant to create automation and equity, algorithms are sometimes hurting employees and consumers and hurling businesses into lawsuits.

We are always choosing. It's complicated, time consuming, annoying. Algorithms are here to help. Right? They confront complexity and bestow time. Algorithms perform a good chunk of our lifting (light and heavy) and spare us the angst of selecting.[2] They inform governments, configure businesses, conduct wars, optimize diets, facilitate love connections, and accompany us wherever we go. But there's an ominous side.

We were not "today years old" when we learned to use algorithms. Wherever we have systemized behaviors, we have algorithms. We get through traffic, decide what to eat, even tie our shoes algorithmically. We've been spinning algorithms since childhood. During gym class, algorithms picked our teammates. We applied one algorithm on basketball day and another when the game was dodgeball. Different algorithms helped us compute the ideal lab partner. Our algorithms worked. Routinely feeding them race, gender, and cultural stereotypes (one set for gym, another for science) helped reinforce our biases.

Letting the algorithm choose made us fair. We used algorithms blindly, but we determined what they saw. All choosing may be biased, but the consequences resonate adversely for vulnerable individuals and groups. Some teammates and lab partners suspected our motivations. But just like individuals encountering algorithm bias today, most never knew a thing. And if they did, they couldn't prove it.

Right now, algorithms are guiding, deciding, and performing a host of functions. From manufacturing to social media, growing collections of processes are trusted to AI. Algorithms recognize and respond to voices, faces, images, names, conditions, and other data. But algorithms don't write themselves (unless aided by another algorithm).

We teach algorithms how to behave. In turn, they assess people and conditions with biases that might be discerned in person but go unnoticed in the digital zone. Most victims of algorithm bias never know that it happened. There is no body language, eye movement, vocal inflection; no bells, alarms, or alerts. And algorithm justice and watchdog groups are just taking their first steps.

II. WHAT IS ALGORITHM BIAS?

We cannot prevent what we cannot define. Algorithm bias seems a contradiction of terms. The algorithm was introduced as a dispassionate decision maker. When the means intended to remove discrimination begins perpetuating it, we have algorithm bias.

Some sources include the word "error" when defining algorithm bias, but this is misleading. "Error" suggests that something in the code is causing the problem, and code can be corrected. Algorithm bias doesn't work that way. Algorithms privilege and victimize different groups, and outcomes are glaringly unfair. But nothing in the original code directly states: if the applicant is female, and the open position is for a chemical engineer, deprioritize the applicant's resume. Yet somehow, bias enters.

As algorithms assume central roles in our AI, their biased behavior is receiving increasing attention. With more of what we do outsourced to AI, bits of our worst are being reactivated. We know about social media platforms allegedly rigging newsfeed algorithms to behave badly to garner clicks. But algorithms exhibiting bias or misbehaving in medical, law enforcement, legal, or other high-stakes settings can kill.

Biased algorithms are taking their place in education, banking, policing, healthcare, social media, dating, marriages, friendships, dieting, hiring, promotion, transportation, manufacture, insurance, investment, space exploration, and beyond. From birth to death, algorithms expose us to bias, and the old patterns are visible. Knowing how it can damage businesses and societies, we struggle to solve bias. Whether you are a corporate or a community, your number one asset is your people. Bias damages the victims and the perpetrators. It reduces organizations and societies. Enter AI.

Artificial intelligence promises to root out bias. But our technology is only as good as our raising, nurturing, and training of that technology. Informing our children that good people come in a variety of races, nationalities, ethnicities, religions, genders, sexualities, and abilities

2 https://blogs.thomsonreuters.com/legal-uk/2018/07/13/ask-dr-paola-algorithms-and-biases-in-recruitment/.

can be difficult for them to adapt if we only expose them to a homogenous selection of good people. Behaviorally, our children may be inclined to associate human goodness with a narrow set of identifying traits. Algorithm bias follows a similar pattern — good intensions undermined by limited examples. It takes more than we realize to prevent our technology from reflecting our history.

Breaches and cybercrime are only some of the perils posed by our tech. More people are exposed to algorithm bias than to computer viruses or bad actors roaming the digital space. Like malware, the timeline of algorithm bias is said to stretch back to the 1980s.[3] It has spawned sub-terminology like algorithm racism and algorithm sexism.[4] There is no shortage of examples. AI discriminates, and we know how.[5]

It's important to note that algorithm bias is all in the training. We encounter the AI, but bias is never the AI's fault. When a facial recognition algorithm that has been trained with more samples of white people's faces has difficulty recognizing a black person's face, we can't be surprised. Our biological eyes are engineered to recognize faces as such; the same cannot be assumed for our algorithms. Imagine if children from racially homogenous communities grew up incapable of recognizing the faces of other racial groups as faces. These shortcomings are just the beginning.

Businesses rely on AI to debias their hiring processes. Many interventions use algorithms that remove triggering data such as names, gender, schools, affiliations, volunteer activities, addresses, and other information. Surviving this initial phase, workers enter companies where other algorithms perpetuate biases that are rooted in race and gender. These biases result from how the algorithm is trained. Face recognition software develops a default bias based on the images used to train it; workplace algorithms learn from data that is skewed towards numerically dominant groups. These groups may be the majority within the organization, the profession, or the society. They may be predominant in a category where the business is trying to change outcomes. For example, the individuals promoted to partner at a professional services firm may be largely of one race and one gender. The algorithm can "see" this fact and may conclude it is a qualification. Algorithm bias can arise when desired outcomes (more women and minorities at the partner level) collide with the preponderance of data related to existing conditions.

III. HOW DO WE DETECT ALGORITHM BIAS?

We cannot delegate to algorithms and hope for the best. Blatant algorithm bias sometimes makes the news. A faulty facial recognition tool led to an innocent Black man's arrest in Detroit. Pedestrian recognition algorithms that have not "seen" enough examples of people with mobility differences cause vehicles to become disproportionately dangerous to some individuals with disabilities. And there is the notorious decision by ChatGPT that marked Australians as less preferred tenants and quickly went viral. But most of the discrimination carried out by algorithm-fed AI never gets on anyone's radar.

Everything is new, but novelty cannot excuse the bias. Ethical oversight of emerging and experimental tech calls for better "teaching" and testing of these consequential tools. New medications do not enter the market without extensive study and testing. After they are launched, sustained scrutiny and reporting is applied, and if the new drug is deemed harmful, it is pulled from the market. Algorithms have a similar broad impact, sometimes involving life and death. Rigorous testing could identify algorithm bias, before and after launch. One algorithm can have various outcomes for different groups and demographics. Bias might be detected if an algorithm is tested repeatedly by diverse parties.

Measuring the bias potential of an algorithm can be achieved through the application of processes that resemble the development of drugs and vaccines. The equivalent of clinical trials might involve highly diverse and inclusive groups of people engaging the algorithm during test phases. Comparing outcomes for these groups could identify and measure bias early in the design process. After they have been tested, adjusted, and released, algorithms must be monitored. An approach analogous to employee feedback and review could be applied to the algorithm to assess and modify the algorithm's evolving thinking and behavior. Afterall, the algorithm is a digital employee. Launching the algorithm and just leaving it to do its thing provides autonomy without oversight.

---

3 Robert P. Bartlett et al. "Algorithmic Discrimination and Input Accountability under the Civil Rights Acts." Available at SSRN 3674665 (2020); Joy Buolamwini and Timnit Gebru."Gender shades: Intersectional accuracy disparities in commercial gender classification." Proceedings of Machine Learning Research: 81:1-15, 2018.

4 Hardesty, Larry. "Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems." MIT News, February 11, 2018. Available at http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 (last accessed April 1, 2023). These companies were selected because they provided gender classification features in their software and the code was publicly available for testing.

5 Hadhazy, Adam. "Biased Bots: Artificial-Intelligence Systems Echo Human Prejudices." Princeton University, April 18, 2017. Available at https://www.princeton.edu/news/2017/04/18/biased-bots-artificial-intelligence-systems-echo-human-prejudices (last accessed April 2, 2023).

IV. WHAT IS CAUSING BIAS?

In their intelligence, algorithms look for data with which to make decisions and build skills. Light or dark skin becomes part of the algorithm's definition of a face.[6] It delivers a Yes or No to the AI in a way that resonates with bygone ideas of who is human and seen and who is not. Did the algorithm arrive at this conclusion without assistance? Or was it supplied with default perspectives which allowed it to mimic long dismissed ideas?

Without exposure or practice, algorithms are deficient. When deployed into real situations, shortcomings become bias. In the workplace, small items of language like gendered pronouns can condition the AI as it seeks to build standards. Unless algorithms are fed diverse data, they are likely to digitize discriminatory default behaviors. Employment, housing, banking, credit and finance, education, and healthcare represent areas where algorithms have injected bias. When we look at the specific environments where algorithm bias crops up, we realized we are observing detectable, measurable phenomena. The biases attributed to algorithms are neither novel nor unique. Bias native to a given industry now plague that industry's algorithms and AI. Biases from banking distort banking algorithms.

A. Algorithm Bias in Employment

Without comprehensive input, algorithms trend toward old biases. Hiring and management algorithms promised to level the playing field for everyone, but encountering more males in engineering candidate pools teaches an algorithm to normalize male dominance in the profession and to use gender as a shortcut when making decisions.

B. Algorithm Bias in Housing

Algorithms across the housing sector help identify high-risk renters or guide builders and developers, but algorithms notoriously interact with complex datasets to make decisions that disadvantage black and brown people. CoreLogic, the company behind CrimSAFE, has been sued because its algorithm is alleged to disproportionately exclude Black and Latino applicants with criminal records.

C. Algorithm Bias in Banking, Credit, and Finance

Discrimination against people of color seeking mortgages and business loans is well documented. Businesses do not purposely use algorithms to perpetuate bias, but things go wrong. A UC Berkeley study showed that equally qualified Black and Brown borrowers pay financial institutions $765 million more per year than White borrowers.

D. Algorithm Bias in Education

When AI becomes the professor or grader, old biases enter. When making decisions about grades or incidents of plagiarism, algorithms show bias. A grade-estimating algorithm used to determine final grades for schools disrupted by COVID-19 favored students from elite private schools and those who were more socioeconomically advantaged.

E. Algorithm Bias in Healthcare

The racial wealth gap, cost of healthcare, and insurance coverage converge to make algorithms discriminatory. Organizations have labored to create diversity among their staff only to have algorithms regress their thinking by several decades. Algorithm bias causes providers to deliver better care to wealthier patients and make decisions that are disadvantageous to people of color.

V. ADDRESSING THE PROBLEM IN ITS INFANCY

AI is not responsible for fixing algorithm bias. Around the globe, many people grow from birth to old age having seen mostly or solely people of their own color or ethnicity. Yet they recognize other faces, and by extension, other humanity. The child or adult who does not recognize a face of different color as a human face might make good sci-fi. Unfortunately, it has become science reality.

AI does not have the data stores of history or genetics. As such, it will extrapolate from our other inputs (or lack of inputs). Algorithms can see whom we include and whom we exclude. It assumes that we like and favor the people who show up most frequently

6 Turner Lee, Nicol. Detecting racial bias in algorithms and machine learning. Journal of Information, Communication and Ethics in Society 2018, Vol. 16 Issue 3, pp. 252-260. Available at https://doi.org/10.1108/JICES-06-2018-0056/ (last accessed April 2, 2023).

on our guest lists. And that's fair. If it finds out that men are more often engineers in our world, it may build a shortcut and use it during hiring or promotion.

Blaming the algorithm transfers responsibility and perpetuates the problem. Wherever algorithms are biased, people did that. People are naturally reluctant to admit they are biased. Most business leaders will claim to have a network of colleagues and contacts that is diverse and inclusive. Close analysis usually reveals that their networks are considerably less diverse than they thought. When designing algorithms, people may believe they are inclusive, but the exposure to default standards teaches an algorithm to adopt behaviors from an untransformed world. It may not matter that the algorithm was devised by someone from a community that is the target of bias. Algorithms have to be directly taught that patterns and events in recent human history have configured the numbers in a certain manner, but it is the job of the algorithm to outperform the defaults and improve outcomes for underserved and underrepresented individuals. Our algorithms have to be reminded that both men and women are engineers. Even if you (the algorithm) see numbers that suggest it is more likely for men to occupy this role, don't use gender as a decision shortcut. Humans must be smart enough to create AI that can recognize and neutralize our bias. We need our algorithms to outperform our defaults.

Fortunately, algorithm bias is still in its infancy. As businesses rollback diversity, equity, and inclusion ("DEI") efforts, it may be left to algorithms to establish parity and close the gaps in the workplace and everyday life. These algorithms must be free of the default biases that we used to design organizations and societies.

Algorithm bias can be fixed. Left unchecked, many of our AI tools will evolve behaviors and methods that profile individuals and groups. Dialogue and innovation hold significant promise when it comes to correcting this dangerous trend. As social media and the world itself takes its first steps in the metaverse, algorithms find new prominence. Managing and eliminating algorithm bias in the virtual, augmented, and mixed realities of the metaverse may be critical to overall success. Algorithm bias will trigger multiple violations of civil rights and other legislation. This issue can be fixed proactively by businesses or addressed through the legal system.

VI. BEST PRACTICES

Left on their own to reach for data and formulate norms, algorithms routinely slip into bias. A series of best practices can help organizations and decision makers get out ahead.

1. Remain open to the possibilities of bias in the algorithms, understand how it might arise, and respond to algorithm bias in a timely manner.
2. Develop standards for determining if algorithm bias is present and disseminate methods of detection.
3. Devise algorithm bias response processes that stipulate what is to be done for a given type of bias incident. Consistently adhere to established protocols.
4. Use diverse teams to formulate, teach, and test algorithms. Let inclusive input and exhaustive testing identify discriminatory tendencies early on.
5. Build and continuously add to a set of terms or criteria that testers can use to determine if an algorithm is biased. Customize terms and criteria for different groups.
6. Ensure that design within algorithms create and sustain equity for all groups.
7. Expose algorithms to exhaustively inclusive datasets as part of a standard development process. To sustain fairness, algorithms must be more diverse than most companies.
8. Among all staff, create a level of psychological safety that lets these individuals feel comfortable voicing their observations regarding the algorithms.
9. Continuously monitor and review the AI as one would an employee. Modify the algorithm based on feedback.
10. Help educate customers regarding the possibility of algorithm bias. Create channels through which they can offer feedback, similar to customer service reporting or a survey tool.
11. Be transparent about how your algorithms work or are supposed to work so that users can begin to assess neutrality or bias in the algorithm and AI.
12. Set a standard minimum level of competence for each algorithm, treating the algorithm like a worker. Determine what is the least level of performance, what are the indicators, and how often will the algorithm undergo a performance review?

As we inadvertently impart active and atavistic discriminatory practices to our AI, algorithm bias shows us who and how we are (and were). Fixing the problem may require remembering all the ways we introduce bias in the first place. We want to win, and that drive can make bias attractive.

We know why we picked a certain person for our basketball team but never considered her for lab partner. It may have turned out that she aced the course. The classmate with the strong accent turned out to be a champion speller. Our biases fail and we are left to rethink. We can let our algorithms know that bias happens, that it is not desired, and teach them not to slip into bias. Screening and retraining algorithms can help a business prevent bias and avoid class actions and other consequences.

People don't question algorithms. It may be in the interest of businesses to prompt consumers to be aware and to offer feedback. As instances of algorithm bias begin to take up more space in news and media, the litigation will follow. The justice department is already settling cases of algorithm bias, and it's only a matter of time before the civil rights implications reach the courts and disrupt vulnerable businesses.

CPI COMPETITION POLICY® INTERNATIONAL

CPI Subscriptions

CPI reaches more than 35,000 readers in over 150 countries every day. Our online library houses over 23,000 papers, articles and interviews.

Visit competitionpolicyinternational.com today to see our available plans and join CPI's global community of antitrust experts.