Mitigating Bias in Artificial Intelligence

An Equity Fluent Leadership Playbook

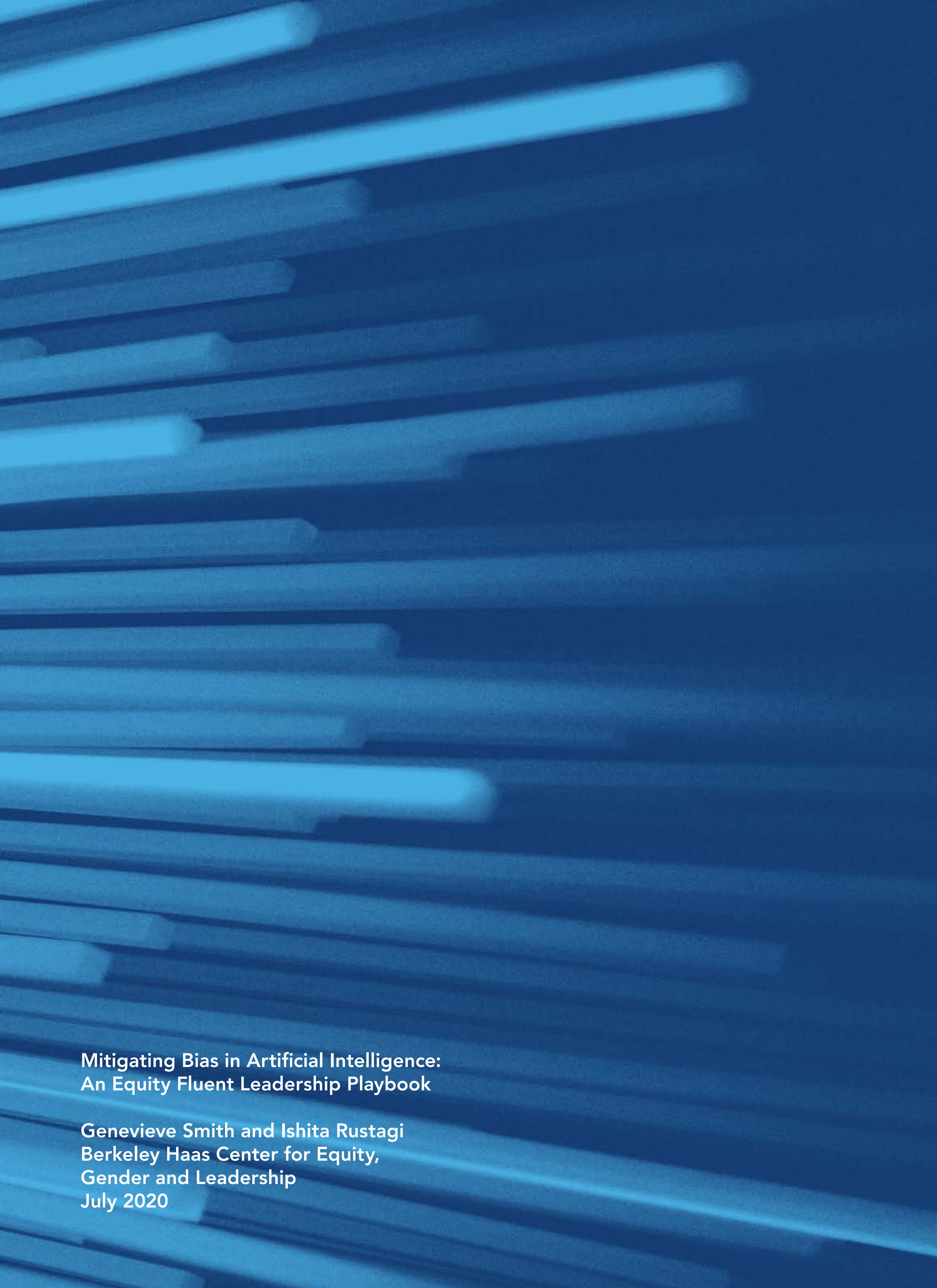
A playbook for business leaders who build & use AI to unlock value responsibly & equitably.

Berkeley Haas

egal

The Center for Equity, Gender and Leadership at the Haas School of Business (University of California, Berkeley) is dedicated to educating equity fluent leaders to ignite and accelerate change. Equity Fluent Leaders understand the value of different lived experiences and courageously use their power to address barriers, increase access, and drive change for positive impact. Equity Fluent Leadership™ (EFL) Playbooks deliver strategies and tools for business leaders to advance diversity, equity, and inclusion. The Playbooks serve as a bridge between academia and industry, highlighting and translating research into practitioner-oriented strategies.

This Playbook was authored by Genevieve Smith and Ishita Rustagi. The Playbook drew from interviews with the following experts who provided invaluable insights: Dr. Adair Morse (Berkeley Haas), Beena Ammanath (Humans for AI), Dr. Cathryn Carson (UC Berkeley), Cansu Canca (AI Ethics Lab), Caroline Jeanmaire (Center for Human-Compatible AI, UC Berkeley), Dr. Gina Neff (Oxford Internet Institute, University of Oxford), Dr. James Zou (Stanford University), Dr. Londa Schiebinger (Gendered Innovations in Science, Health & Medicine, Engineering, and Environment, Stanford University), Dr. Paul Gertler (Berkeley Haas), Dr. Stuart Russell (Computer Science, UC Berkeley), and Sarah West (AI Now Institute). The playbook benefited from helpful feedback from Alberto Melgoza (Google), Abigail Mackey (Berkeley Haas EGAL), Dominique Wimmer (Google), Francesca LeBaron (Berkeley Haas EGAL), Jamie Ellenbogen (Google), Jennifer Wells (Berkeley Haas EGAL), Jessa Deutsch (BCG), Jesse Kaiser (BCG), Katia Walsh (Levi Strauss & Co.), Dr. Kellie McElhaney (Berkeley Haas EGAL), Paul Nicholas (Google), Paul Spurzem, Sarah Allen (Google), and Teresa Escrig (Microsoft). The playbook also benefited from a Working Group including Nitin Kohli (UC Berkeley School of Information) and Jill Finlayson (Women in Technology Initiative, UC Berkeley), as well as contributions from Aishwarya Rane and Matt McGee (UC Berkeley).



Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook

Genevieve Smith and Ishita Rustagi
Berkeley Haas Center for Equity,
Gender and Leadership

July 2020

What is this playbook?

Mitigating Bias in AI: An Equity Fluent Leadership Playbook provides business leaders with key information on bias in AI (including a Bias in AI Map breaking down how and why bias exists) and seven strategic plays to mitigate bias.

The playbook focuses on bias particularly in AI systems that use machine learning.

Who is this playbook for?

You are a CEO, a board member, an information / data / technology officer, a department head, a responsible AI lead, a project manager… No matter where you fall in your organizational chart, you see yourself as a leader who is eager to respond to the bigger picture opportunities and risks of AI for your customers, shareholders, and other stakeholders.

Why use it?

This playbook will help you mitigate bias in AI to unlock value responsibly and equitably. By using this playbook, you will be able to understand why bias exists in AI systems and its impacts, beware of challenges to address bias, and execute seven strategic plays.

How to use this playbook?

The Playbook includes a "Snapshot" that outlines top-line information on bias in AI, strategic plays to address bias, and steps to put them into action. It also includes a "Deeper Dive" that delves deeper into bias in AI, impacts for businesses and society from biased AI, and challenges for businesses to address it. If you are an AI practitioner, not familiar with bias in AI, somewhat familiar with bias in AI or tend to see bias in AI as more of a technical issue – we recommend exploring the "Deeper Dive".

Guides for each of the plays – including how-to information, mini case studies of leading businesses, and tools – can be found separately on our **Playbook site**.

How was this playbook developed?

The Playbook was developed through leading expert interviews; a review of the literature across various disciplines such as engineering, sociology, data science, anthropology, philosophy, and more; as well as collection and analysis of bias in AI examples across industries and AI applications. It was prototyped and iterated with businesses and business leaders.

Contents

Foreword

"Another batch of candidates that are almost all white men? This is curious."

Anita and her other colleagues work on the hiring team at a Bay area tech firm and were not phased the first time that the top candidates recommended for interviews were white men – tech companies are, after all, predominantly filled with white, male employees. But as the trend continued, Anita and her team took pause. The company had just started using an artificial intelligence (AI) system that helped her team save countless hours by working through piles of applications to identify the top candidates to move onto the interviewing stage.

When Anita approached the developers highlighting this trend they pushed back at first. The AI system – using machine learning – had been trained on data from the company's current employees, as well as past applicants with the purpose of identifying the best candidates for each position. It had been designed to be "gender-blind" and "race-blind" so it should be unbiased – or so they thought. But digging into it further, the developers (who, reflecting the technical employee base at the company, were predominantly white men) found that the AI system did indeed have a bias – candidates with resumes including words associated with women were penalized. The AI system had learned to be biased and they couldn't figure out how to "de-bias" it. Eventually, the team was disbanded and the originally promising system was scrapped.

Anita and her firm's story is not unique and one illustration of bias in AI systems and how it can be a silent killer for firms. Bias can creep in – through the data and throughout the development and evaluation of algorithms that compose the AI system. It is related to and reinforced by those who are designing, managing, and using AI systems. Bias in AI is a larger business issue that requires various actions and efforts that can and should be overseen by business leaders before it is too late, immense risk is realized, and opportunity is lost.

Currently, organizations don't have the pieces in place to successfully mitigate bias in AI. But with AI increasingly being deployed within and across businesses to inform decisions affecting people's lives, there is too much at stake – for individuals, for businesses, and for society more broadly.

Much has been written about bias in AI with largely technical guidance, but doesn't always incorporate academic literature across disciplines and speak to the larger business solutions and opportunities. We developed this Playbook to address the gap between knowledge and action for business leaders and recognizing that AI is here to stay – but new approaches are needed.

The Snapshot

Artificial intelligence (AI) makes it possible to automate judgments that were previously made by individuals or teams of people. Using technical frameworks such as machine learning, AI systems make decisions and predictions from data about people and objects related to them.

AI represents the largest economic opportunity of our lifetime – estimated to contribute $15.7 trillion to the global economy by 2030 according to PwC research.[1] Businesses leaders at IBM anticipate adoption of AI in the corporate world to explode up to 90% in the next 18-24 months.[2]

AI is increasingly employed to make decisions affecting most aspects of our lives, particularly as digital transformation is accelerating in the face of COVID-19. AI informs who receives an interview for a job, whether someone will be offered credit, which products are advertised to which consumers, as well as how government services and resources are allocated – such as what school children will attend, who gets welfare and how much, which neighborhoods are targeted as "high risk" for crime, and more. For emergency response to COVID-19, AI is helping identify the virus, inform allocation of resources to patients in hospitals, and support contact tracing. Use of AI in predictions and decision-making can reduce human subjectivity, but it can also embed biases resulting in inaccurate and/or discriminatory predictions and outputs for certain subsets of the population.

Harnessing the transformative potential of AI requires addressing these biases, which pose immense risk to business and society. As developers, users and managers of AI systems, businesses play a central role in leading the charge while decisions of business leaders are of historic consequence.

The goal is not to fully "de-bias" AI – this is not achievable. Bias in AI isn't simply technical and can't be solved with technical solutions alone. Addressing bias in AI requires assessing the playing field more broadly. It requires seeing the big picture – where different business roles and players fit in, how they pass and play together, where the ball is coming from and where it should go. This is why addressing bias in AI is an issue for business leaders – for the coaches in governance and captains within departments or teams. Addressing bias in AI requires business leaders to see, direct and navigate strategies.

The ultimate goal is to mitigate bias in AI to unlock value responsibly and equitably. By using this playbook, you will be able to understand why bias exists in AI systems and its impacts, beware of challenges to address bias and execute strategic plays.

AI could contribute

$15.7 trillion

to the global economy by **2030**

This playbook focuses on machine learning AI systems (which we refer to in this playbook as simply 'AI systems'). Machine learning is a common and popular subset of AI used for predictions and decision-making, but has clear limitations and issues related to bias. If you are interested in machine learning AI, this playbook is for you – read on.

How does biased AI impact business?

Addressing bias in AI is not only the right thing, but the smart thing for business – and the stakes for business leaders are high. Biased AI systems are those that result in inaccurate and/or discriminatory predictions and outputs for certain subsets of the population. Biased AI systems can unfairly allocate opportunities, resources or information; infringe on civil liberties; pose a detriment to the safety of individuals; fail to provide the same quality of service to some people as others; and negatively impact a person's wellbeing such as by being derogatory or offensive.

These issues cost businesses by negatively impacting their reputation, consumers' trust, and future market opportunities. Tech companies recognize this risk: Microsoft flagged reputational harm or liability due to biased AI systems as a risk to their business in a report to the US Securities and Exchange Commission.[3]

AI systems found to be biased may be scrapped or need significant changes, resulting in high costs in terms of employee time and other invested resources. For example, in 2018, Amazon recalled its AI-driven hiring tool designed to mechanize the search for top talent when it was found to be biased against women, penalizing candidates whose resumes included the word "women's".[4] Biased AI can also cause internal conflicts and employee demand for more ethical practices.

At a larger societal level, AI systems can solidify and amplify societal discrimination, while discriminatory resource allocation can lead to inefficiencies and losses in the economy and markets. Recognizing these issues, governments are pursuing regulation and legislation. Companies that don't make addressing bias in AI a priority may be liable to incur large penalty fees.

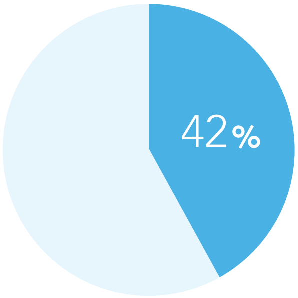


Figure 1. A 2019 DataRobot report found that **42% of organizations currently using / producing AI systems are "very" to "extremely" concerned** about the reputational damage that media coverage of biased AI can cause.[13]

BY TACKLING BIAS IN AI SYSTEMS THROUGHOUT THE DEVELOPMENT AND MANAGEMENT OF THESE SYSTEMS, BUSINESSES CAN...

- Mitigate risk
- Maintain strong brand reputation
- Have a superior value proposition
- Stay ahead of forthcoming legislation
- Be a competitive leader in the fast-paced industry

Why are AI systems biased?

At a high-level, AI systems are biased because they are human creations. They are classification technologies and are products of the context in which they are created.[5] Unsurprisingly then, they often mirror society.[6]

It matters who develops AI systems. The perspectives and knowledge of those who develop AI systems are integrated into them, while the values and priorities of managers and business leaders impact an organization and the products it develops.

Tech companies and labs developing many large-scale AI systems tend to be mainly male and white. The share of women in computing today is 26% – lower than it was in 1960.[7] Almost half of women who go into tech leave the field, which is more than double the percentage of men who depart.[8] Despite being half the population, less than one fifth of AI researchers or professors are women (see Figure 2). Racial diversity also lacks. As seen in Figure 3, at Microsoft, 4.5% of employees are Black and 6.3% are Hispanic/Latinx.[9] These numbers are similar at other tech firms. Beyond demographic diversity, in many cases AI systems are not designed with relevant domain experts, are not adapted to the particular context in which they are used and are not informed by end users.

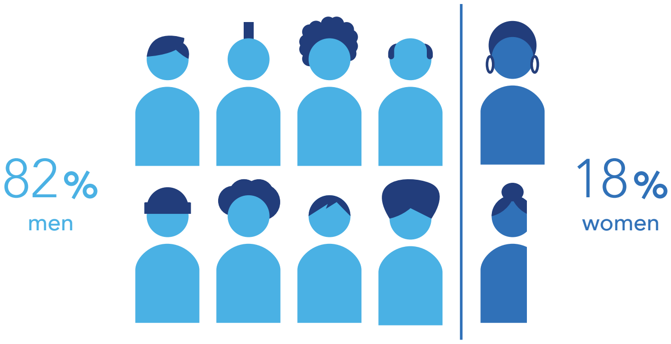


Figure 2. Only 18% of researchers at leading AI conferences are women

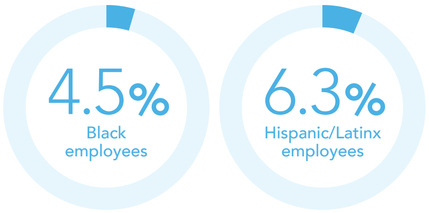
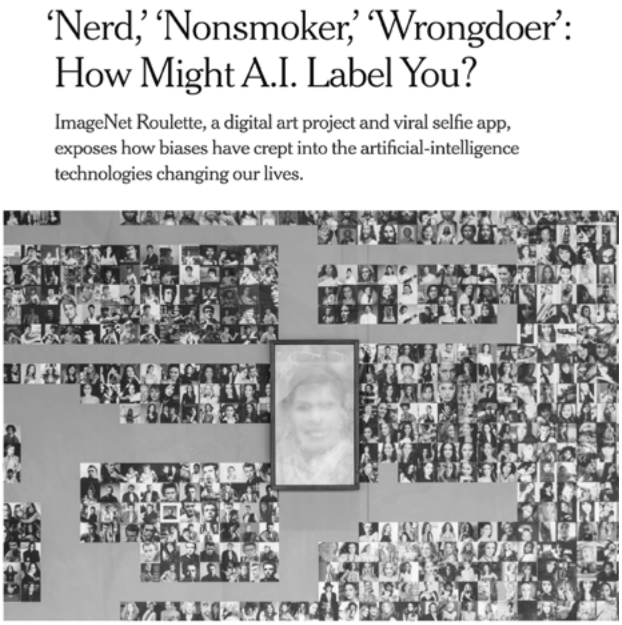


Figure 3. At a leading US tech company only 4.5% of employees are Black, and only 6.3% are Hispanic / Latinx

More specifically, bias can enter in the development and use of a machine learning AI system. It can enter in the generation, collection, labeling and management of data that the algorithm learns from; as well as the design and evaluation of algorithms. It happens largely unknowingly and despite noble intentions. Biased AI results in inaccurate predictions and/or discriminatory outputs and predictions that then pose immense hazards for individuals and business (see our **Bias in AI Map,** Figure 4). We briefly describe how bias enters AI systems here, while a detailed breakdown of the Map, including specific

pathways that bias can enter in datasets and algorithms, as well as use of AI systems, is found in the Deeper Dive.

Let's start with data. The dataset(s) used to train an algorithm is critical–AI systems learn to make decisions based on these training datasets – and there are various points where bias can enter. Vast amounts of data points are **generated** by virtue of individuals' day-to-day activities (e.g., consumer behavior, health conditions) and data points are **collected** through various platforms, technological or otherwise. Data is assumed to accurately reflect the world, but there are significant data gaps (including little or no data coming from particular communities[10]) and data is rife with racial, economic and gender biases.[11] In addition, human influence cannot be eliminated from data. In many cases, humans decide what, where and how data is collected and categorized, as well as parameters for a **dataset**. Data is also **labeled**, which can be subjective. Data collected in the past preserves and reflects that past.[12]

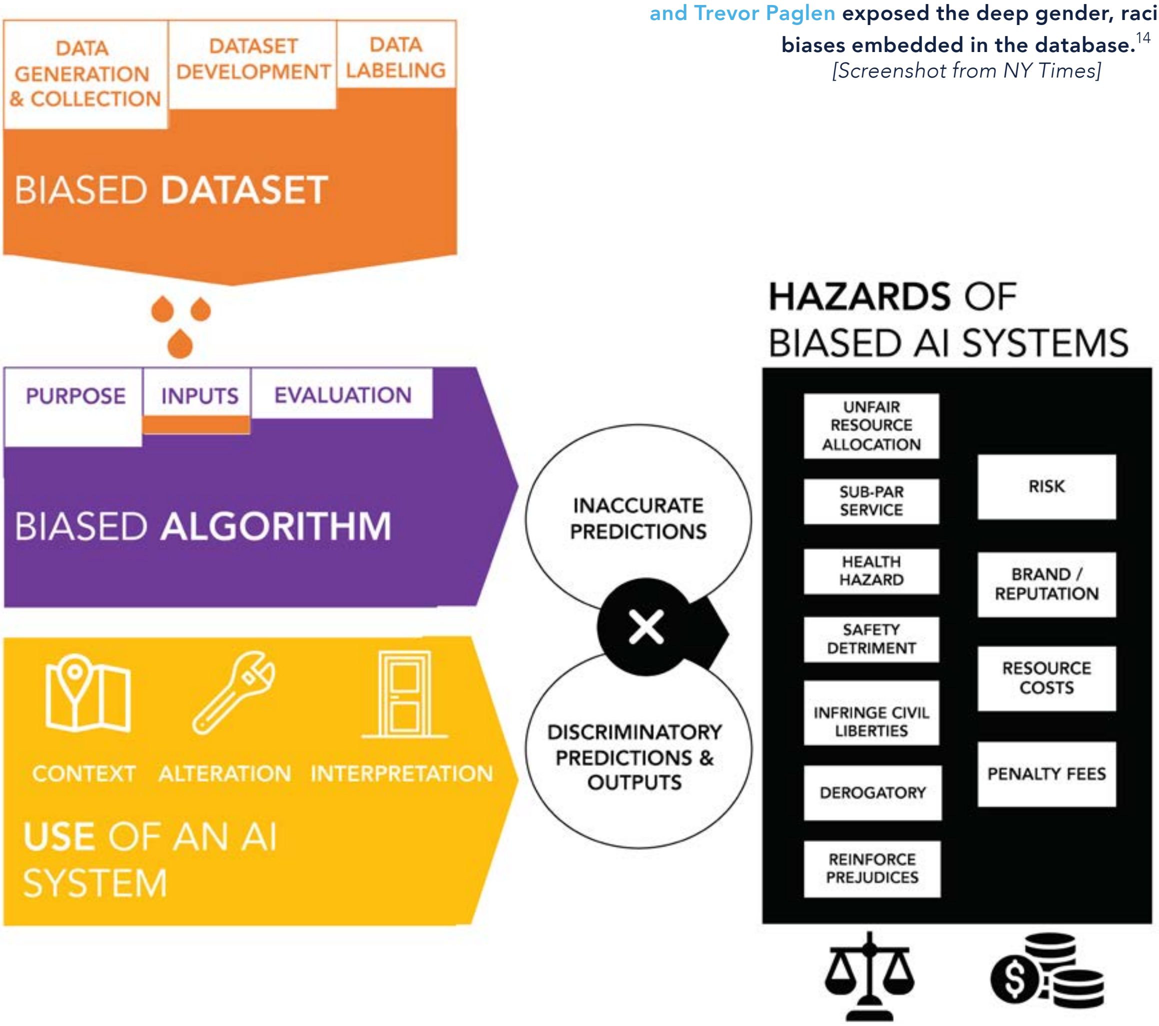


'Nerd,' 'Nonsmoker,' 'Wrongdoer': How Might A.I. Label You?

ImageNet Roulette, a digital art project and viral selfie app, exposes how biases have crept into the artificial-intelligence technologies changing our lives.

Figure 5. ImageNet, developed by researchers at Stanford, is a widely used database with millions of images that computer vision AI technologies learn from. Historically, images mainly included photos from the US, and various photos were classified problematically – including labels like "nerd" and "slut". ImageNet Roulette, an art project by Kate Crawford and Trevor Paglen exposed the deep gender, racial and other biases embedded in the database.[14]
[Screenshot from NY Times]

FIGURE 4. BIAS IN AI MAP



A biased dataset can be unrepresentative of society by over or under-representing certain identities in a particular context. Biased datasets can also be accurate but representative of an unjust society. In this case, they reflect biases against certain groups that are reflective of real discrimination that the particular group(s) face.

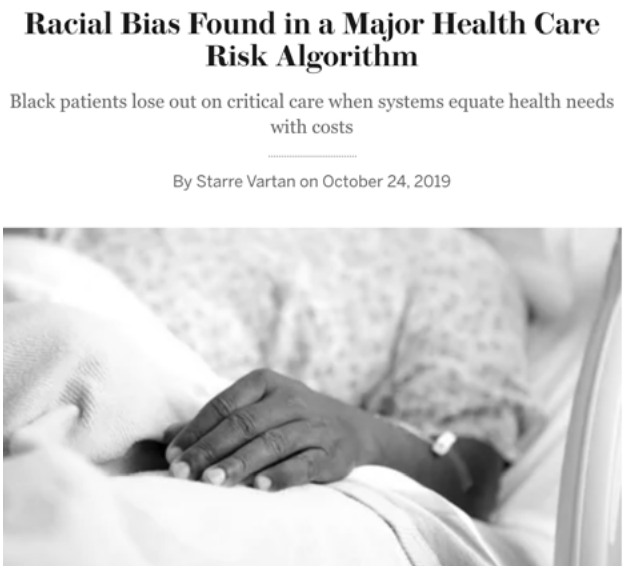
Bias can enter an algorithm at various points too. It can creep in when defining the **purpose** of an AI model and the constraints it operates under. Bias can enter when selecting the **inputs** the algorithm should consider to find patterns and draw conclusions. This includes selection of datasets that the algorithm should learn from and selection of proxies / variables.

An algorithm can lead to technically inaccurate predictions for certain subsets of the population. This happens particularly if drawing from a dataset in which a certain identity is over- or under-represented or if a dataset is used that is not specific to the algorithm's target population. An algorithm can contribute to discriminatory outputs irrespective of the quality of training data used, depending on how it is **evaluated.**

Beyond a biased algorithm(s), AI systems can result in discriminatory outcomes for certain individuals or populations based on how they are used. There is potential for inaccurate predictions and bias if an AI system is used in a different context or for a different population from which it was originally developed or if it is applied for different use cases from which it was originally developed / operationalized. AI systems can be used or altered by organizations or individuals in ways that can be deemed as discriminatory for certain populations. This can be due to bad actors getting ahold of and using the technology. In other cases, it may be less overt and subject to debates over fairness. Lastly, for AI systems that support human decision making, how individuals interpret the machine's outputs can be informed by one's own lived experience and allow for bias.

Use of the AI system resulting in discriminatory outcomes is not the main focus of this playbook, but important to acknowledge.

For more information and examples on the different ways bias can enter AI systems, see the Deeper Dive.

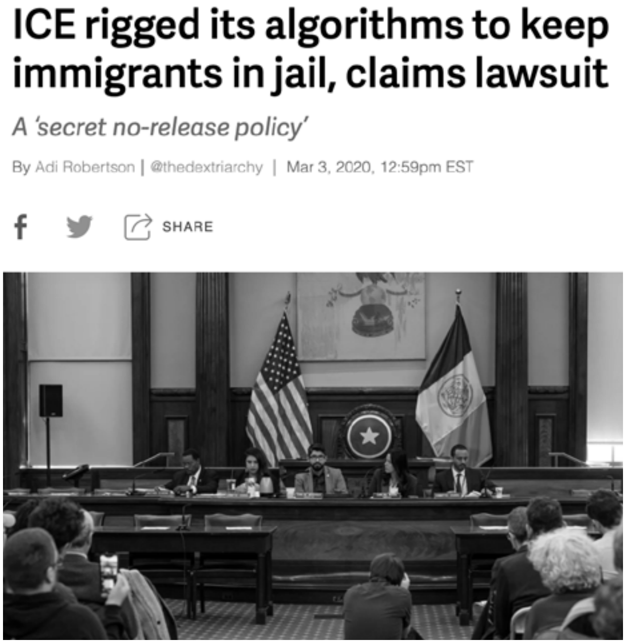


Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

Figure 6. A widely used healthcare algorithm in the US falsely concluded that Black patients are healthier than equally sick White patients. The algorithm uses health costs as a proxy for health needs, but Black patients who have the same level of need spend less money because of unequal access to care, among other reasons.[15]
[Screenshot from Scientific American]



ICE rigged its algorithms to keep immigrants in jail, claims lawsuit

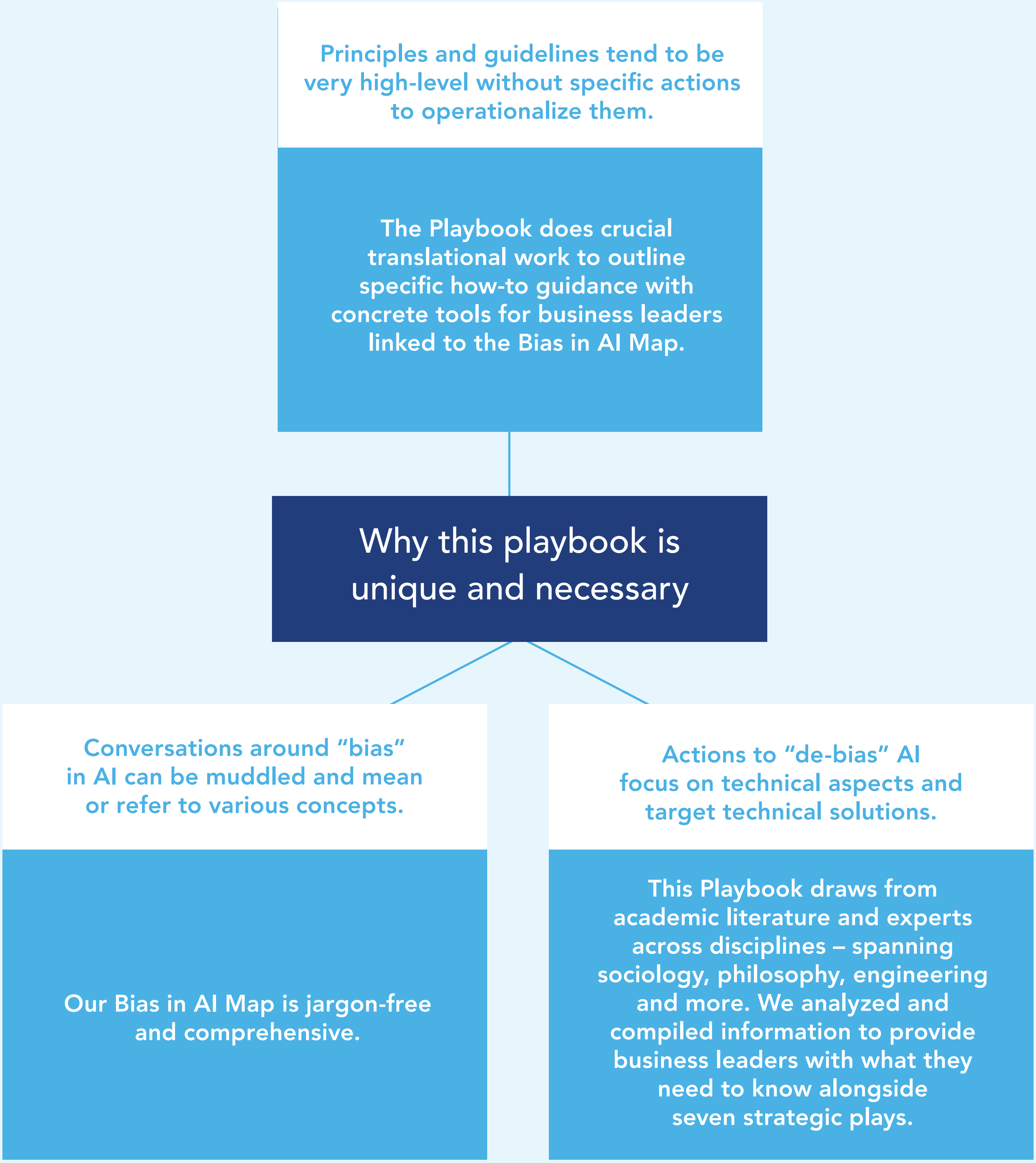
A 'secret no-release policy'

By Adi Robertson | @thedextriarchy | Mar 3, 2020, 12:59pm EST

Figure 7. Early 2020, the New York Civil Liberties Union and Bronx Defenders filed a lawsuit claiming the Immigration and Customs Enforcement (ICE) rigged an AI system to create a "secret no-release policy" for people suspected of breaking immigration laws.
[Screenshot from The Verge]

What action is being taken & where does the playbook fit in?

Action to understand and mitigate bias in AI is being taken by various stakeholders – spanning companies, academia, government, multilateral institutions, non-governmental organizations (NGOs), and even the Roman Catholic Church.[16] Along with this is a proliferation of principles, guidelines, and pushes for industry-wide responsible AI regulations and practices. Many companies are engaging in partnerships, leading efforts, and advancing responsible AI internally. Despite this, gaps remain for business leaders - which is where this playbook comes in.



Principles and guidelines tend to be very high-level without specific actions to operationalize them.

The Playbook does crucial translational work to outline specific how-to guidance with concrete tools for business leaders linked to the Bias in AI Map.

Why this playbook is unique and necessary

Conversations around "bias" in AI can be muddled and mean or refer to various concepts.

Our Bias in AI Map is jargon-free and comprehensive.

Actions to "de-bias" AI focus on technical aspects and target technical solutions.

This Playbook draws from academic literature and experts across disciplines – spanning sociology, philosophy, engineering and more. We analyzed and compiled information to provide business leaders with what they need to know alongside seven strategic plays.

What are the strategic plays for business leaders?

Strategic plays for business leaders to mitigate bias in AI span three buckets - teams, AI model, and corporate governance and leadership.

The following pages include questions that each strategic play addresses. Strategic plays should be part of your longer term game plan. Some plays also have 'quick wins', which are brief, accompanying resources that can be implemented this quarter and have concrete, immediate benefits. More information on the quick wins and strategic plays - with how-to-guidance, mini cases of leading businesses, and tools - can be found separately on our **Playbook site***.*

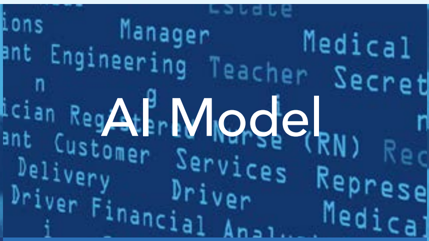
Teams

1

Enable diverse and multi-disciplinary teams working on algorithms and AI systems.

2

Promote a culture of ethics and responsibility related to AI.



AI Model

3

Practice responsible dataset development.

4

Establish policies and practices that enable responsible algorithm development.

Corporate governance & leadership

5

Establish corporate governance for responsible AI and end-to-end internal policies to mitigate bias.

6

Engage corporate social responsibility (CSR) to advance responsible / ethical AI and larger systems change.

7

Use your voice and influence to advance industry change and regulations for responsible AI.

Teams

1. Enable diverse and multi-disciplinary teams working on algorithms and AI systems.

Have diverse teams researching, developing, operationalizing and managing algorithms and AI systems. Even with diverse teams, we all have blind spots and the onus cannot be simply placed on diverse individuals to identify and mitigate biases. Engaging individuals in the social sciences and humanities – as well as domain experts that understand the particular domain the AI system is meant to operate in – is important.

QUESTIONS:

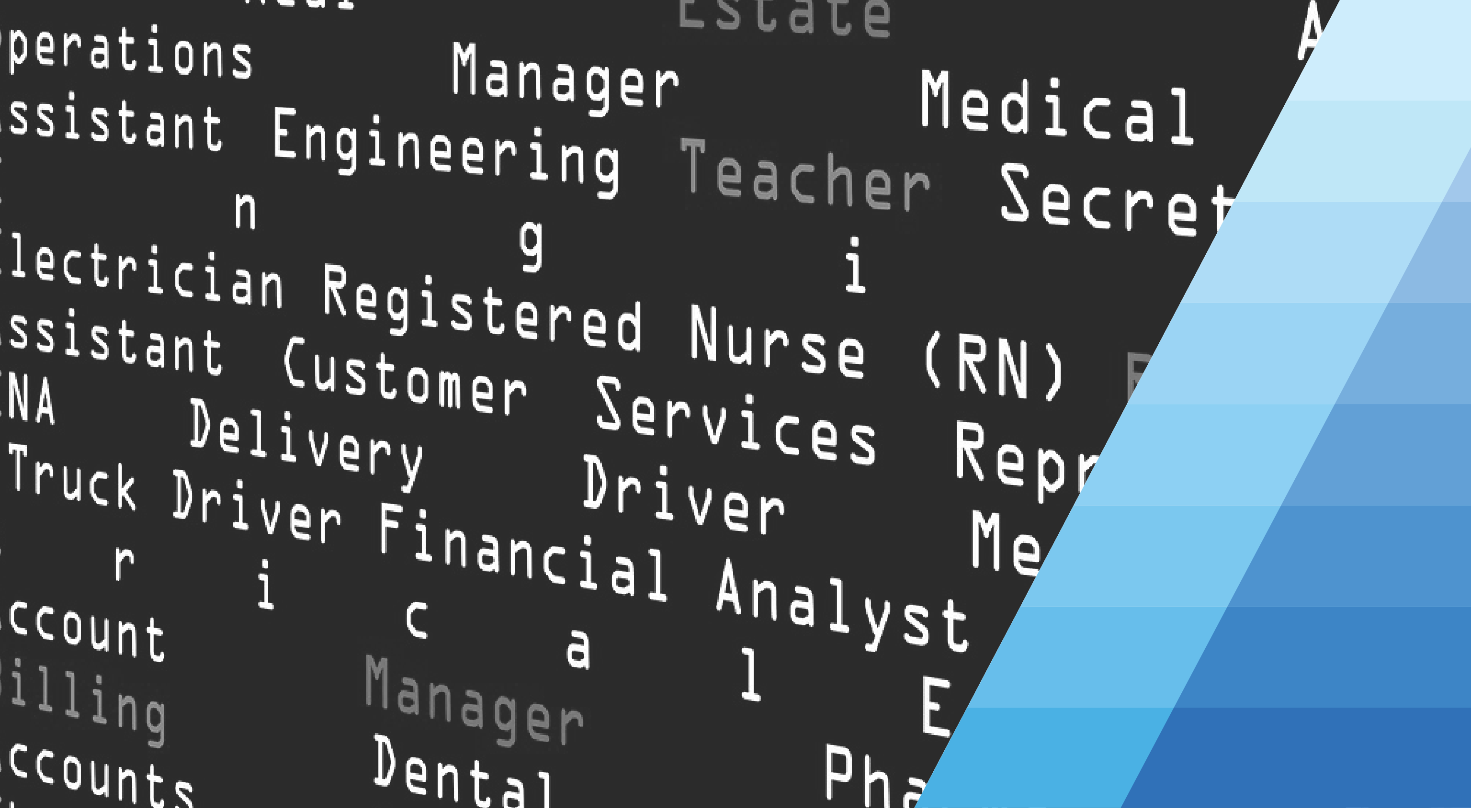
- Is diversity a core priority for leadership?
- Are institutional policies and practices developed and in place to explicitly support women and people of color?
- Does the organizational culture promote diversity, equity and inclusion (DEI)?
- Are teams developing, managing and using AI systems multi-disciplinary?
- Does the company invest in enhancing diversity in data science and engineering?
- Does the company invest in research to understand DEI in tech and AI?

2. Promote a culture of ethics and responsibility related to AI.

Enable a culture that empowers and encourages employees to prioritize equity considerations at every step of the algorithm development process. In line with the understanding that completely de-biasing AI may not be feasible, organizations should uphold a standard of explainability around the workings of their models, as well as transparency around potential shortcomings / pitfalls.

QUESTIONS:

- Are staff expected and incentivized to flag ethical issues and promote responsible AI? Is this priority reflected in performance reviews, as well as individual and team goals?
- Are staff trained on ethical considerations as they relate to AI, bias and fairness?
- Are staff and/or contractors labeling data trained on language related to equity and inclusion?
- Is it expected that staff / teams incorporate explainability and transparency around shortcomings and pitfalls of AI systems?


Estate
perations Manager Medical
ssistant Engineering Teacher Secret
lectrician Registered Nurse (RN)
ssistant Customer Services
NA Delivery Repr
Truck Driver Driver Me
Financial Analyst
ccount
illing Manager
ccounts Dental Pha

AI Model

3. Practice responsible dataset development.

Ensure that dataset development is conducted responsibly, with standard checks and balances in place for creating new datasets as well as adapting existing ones. The creation and implementation of such practices requires businesses to be intentional about gathering inclusive data and asking important questions around who is benefiting from the data collected.

QUESTIONS:

- Do teams developing machine learning datasets assess the quality and quantity of data generated and gathered to ensure populations are sufficiently and accurately represented?
- Do teams developing machine learning datasets ensure that existing datasets are not being appropriated for uses they may not be built / suited for?
- Do teams developing machine learning datasets document their provenance, creation, and use?

4. Establish policies and practices that enable responsible algorithm development.

Build practices that check for and actively mitigate bias into every stage of the algorithm development process. This involves equipping teams with ethical frameworks that allow them to prioritize equity while defining their algorithms' objectives, ensuring datasets used are responsibly developed and labeled, and ensuring variables do not disadvantage certain communities.

QUESTIONS:

- Is the development process standardized with tools to identify, document and mitigate the shortcomings and risks of the AI model?
- Does the team consider where and how to integrate human-in-the-loop processes?
- Is the AI system audited – including internal and external AI audits?
- Are there robust feedback mechanisms built into AI systems so users can easily report performance issues they encounter, and (if no way to opt out), have an appeal process to request human review?

Corporate governance & leadership

5. Establish corporate governance for responsible AI and end-to-end internal policies to mitigate bias.

Establish corporate governance for responsible AI and end-to-end internal policies to mitigate bias. AI ethics governance structures is a first step.

QUESTIONS:

- Does the company have clear leadership for responsible AI, such as an AI ethics lead and AI ethics board?
- Does the company have an AI ethics code / principles?
- Do current leadership priorities prioritize efficiency potentially at the cost of ethical and responsible AI?
- Does the company have formal processes (including concrete guidance and accountability structures) to help plan for, identify and mitigate biases in AI systems?

6. Engage corporate social responsibility (CSR) to advance responsible / ethical AI and larger systems change.

Leverage CSR teams, operating under different incentive structures than other parts of the business with less of a priority on efficiency, to advance responsible AI internally. CSR teams can also be deployed to address biases in data; address power dynamics and lack of diversity in AI; and catalyze research and education (for data scientists, engineers and business students) on responsible AI.

QUESTIONS:

- Has the company leveraged the CSR team to advance internal bias mitigation efforts?
- Are CSR efforts aligned with the company's goals and material interests to mitigate bias in AI and support long term systems change?

7. Use your voice and influence to advance industry change and regulations for responsible AI.

A responsible business leader understands that bias in AI is not simply a technical issue and sees the trade-offs related to "fairness" that can be at play. Business leaders can use their voice and influence to support industry change and advance much-needed regulations.

QUESTIONS:

- Are you part of meaningful partnerships with various stakeholders to inform or advocate for policies for responsible AI and approaches in industry?
- Do you contribute to ongoing debates around bias in AI and insist on / support meaningful dialogue among a wider array of stakeholders in algorithmic accountability?
- Does the company fund research to advance knowledge in the space of responsible AI (especially diverse research teams) and prioritize working with other organizations or initiatives that have diverse teams and/or responsible data / AI systems practices?

Putting the plays into action

You are a CEO, a board member, an information / data / technology officer, a department head, a responsible AI lead, a project manager… No matter where you fall in your organizational chart, you see yourself as a leader who is eager to respond to the bigger picture opportunities and risks of AI for your customers, shareholders, and other stakeholders.

… Where to begin?

STEP 1

Get yourself and other internal leaders up to speed.

□ Share this "Snapshot" with other internal leaders to understand (at a high level) why bias in AI systems is a problem for business, how it manifests and what to do.

STEP 2

Understand the nature of the beast.

□ Have your direct reports, project managers and others that have a mandate to advance responsible AI read the "Deeper Dive".

STEP 3

Gather support internally to execute strategic plays.

□ Have you gathered internal support? You know your company and context best, but here are some ideas to gather internal support:

- Highlight the business case for addressing bias in AI that is outlined in this playbook, emphasizing risk mitigation.
- Use examples in your industry and application(s) of AI where mitigating bias unlocked new value, or where ignoring bias led to costly avoidable mistakes. Find relevant examples using our **Bias in AI Examples Tracker** or reading the Deeper Dive.
- Look internally to review any previously known biases or issues with the AI systems you use or develop; take note of potential reputational or legal risks.
- Connect mitigation of bias to values of the company and achieving corporate goals
- To gather support for projects and initiatives: Link / connect the importance of mitigating bias in AI to achieving specific OKRs (Objectives and Key Results)

Put the plays into action.

STEP 4

□ Designate a person to be responsible for each play and ask them the questions highlighted under each play.

□ Schedule a meeting with each play lead for one month from now to discuss the play and their action plan. Each responsible lead can come prepared with:

- Where can the company grow?
- Which of the outlined steps should the company take / how might we customize or adapt them?
- What tools should we use?

□ Get a 'quick win' under your belt. Some plays have 'quick wins' which are brief resources that can be implemented this quarter and have immediate benefits.

- We recommend starting with this 1.5-hour **Case Study on Bias in AI** that can be done over a brown bag lunch (associated with Play #2).

The Deeper Dive

Artificial intelligence (AI) is increasingly being used to make decisions and predictions affecting most aspects of our lives. This includes uses spanning who receives an interview for a job, whether someone will be offered credit, which products are advertised to which consumers, as well as government services and resource allocation – such as what school children will attend, who gets welfare and how much, which neighborhoods are targeted as "high risk" for crime, and more. For emergency response to COVID-19, AI is helping identify the virus, inform allocation of resources to patients in hospitals, and support contact tracing.

AI has incredible potential to make decisions more efficient and cost-effective, while also promoting higher productivity growth in the economy. A PwC survey finds that 85% of CEOs believe that AI will significantly change the way they do business in the next 5 years[17] and PwC research estimates that AI could contribute $15.7 trillion to the global economy by 2030.[18] As COVID-19 continues to deeply and severely impact people, communities and economies in all corners of the world, there is greater reliance on digital technologies than ever before. Innovative technologies using AI systems might help the expected low GDP growth and productivity in the coming years and help speed critical economic recovery from COVID-19.[19] Use of AI in predictions and decision-making can reduce humans' subjectivity, but also embed biases, produce discriminatory outcomes at scale and pose immense risks to business.

The goal is not "de-biasing" AI – this is not achievable. Bias in AI isn't simply technical and can't be solved with technical solutions alone. Addressing bias in AI requires assessing the playing field more broadly. It requires seeing the big picture – where different business roles and players fit in, how they pass and play together, where the ball is coming from and where it should go. This is why addressing bias in AI is an issue for business leaders – for the coaches in governance and captains within departments. Addressing bias in AI requires business leaders to see, direct and navigate strategies. Business leaders who do will keep ahead of rivals and enable AI to reach its unlimited potential.

The ultimate goal is to mitigate bias in AI to unlock value responsibly and equitably. By using this playbook, you will be able to understand why bias exists in AI systems and its impacts, beware of challenges to address bias, and execute strategic plays.

understand

beware

execute

85%

of CEOs

believe that AI will significantly change the way they do business in the next

5 years

AI could contribute

$15.7 trillion

to the global economy by **2030**

Background

What is AI and why is it so commonly used today?

The use of AI continues to grow following progress in machine learning (ML) - particularly deep learning (DL) - and the proliferation of data.

The field of AI can be traced to the 1950s. Alan Turing, a British mathematician, considered how "thinking machines" could reason like humans and developed the Turing test to determine a machine's intelligence. Later, John McCarthy, a professor at MIT, coined the term "artificial intelligence".[20] Today, AI generally refers to "machines that respond to stimulation consistent with traditional responses from humans, given the human capacity for contemplation, judgment and intention."[21]

Basically, AI is the use of a computer to model and/or replicate intelligent behavior. Algorithms (which are mathematical instructions that give instructions to computers) are designed by humans and used in AI systems to make decisions using data.
ML is used in the majority of AI applications. ML – which is made up of a series of algorithms – takes and learns from massive amounts of data to find patterns and make predictions. It performs a function and gets progressively better over time.

DL is a subset of ML (see Figure 1) in which models make their own predictions independent of humans – after the models are created. In DL, algorithms are structured in layers and modeled after the biological neural network of the human brain. DL makes it possible for the computer program to learn on its own.[22] It is called "deep learning" because it has (deep) layers that help it to learn from data. For example, in facial recognition applications, DL algorithms learn to recognize the nose at one level and eyes at another level. Over time, it improves performance.[23]

Machine (and deep) learning can be supervised, unsupervised and reinforced. For supervised

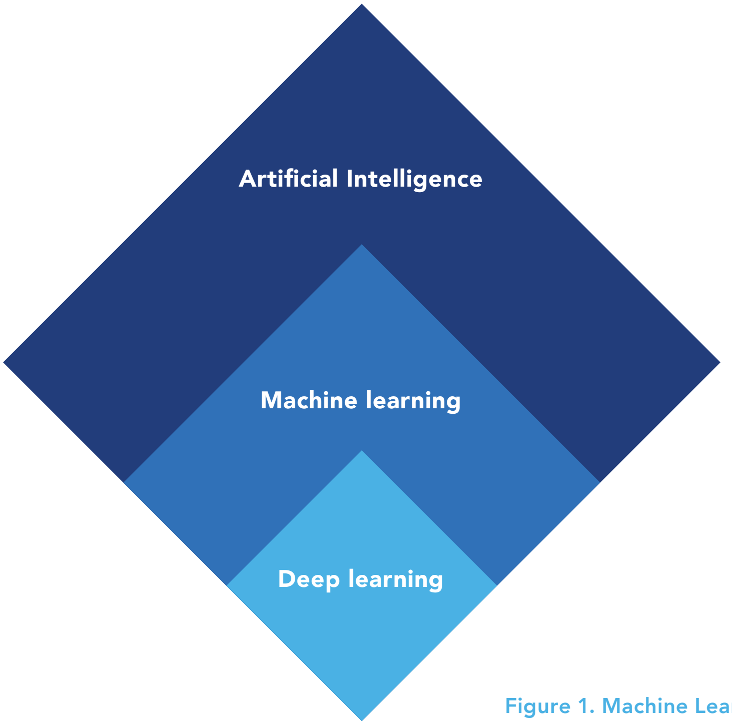


Figure 1. Machine Learning is a subset of Artificial Intelligence. Deep Learning is a subset of Machine Learning.

learning (the most prevalent), data is labeled to tell the algorithm exactly what patterns to look for. For unsupervised learning, data has no labels so the algorithm looks for patterns itself. A reinforcement algorithm learns by trial and error to achieve a certain objective (e.g., AlphaGo is a computer program that learned by this trial and error and eventually defeated a professional Go player).[24]

AI is everywhere today because of recent advancements in ML and DL, as well as massive amounts of data now available. AI relies on data – and large volumes of it. Data encompasses a lot – it can be numbers, words, images, clicks, etc. The more data ML models have, the more accurate they become.[25] Massive infrastructure has made it possible to process, compile and digitize data -- which can then be fed into an ML algorithm. While data has been steadily increasing over time, there was a sharp rise since 2010.[26]

ML/DL models have limitations. They can incorrectly equate correlation identified by the algorithm as causation. Since it is hard to know why algorithms make certain predictions, these types of errors are problematic. Also, by using and learning from data collected at some point in the past to make predictions, ML/DL models project the past into the future. An important question is whether ML / DL is the right technical system to tackle the problem at hand.

This playbook focuses on machine learning AI systems (which we refer to in this playbook as simply 'AI systems'). Machine learning is a common and popular subset of AI, but has clear limitations and issues related to bias. If you are interested in machine learning AI, this playbook is for you – read on.

Investment in AI systems is accelerating, trust is critical

Development and use of AI systems is increasing exponentially. Businesses leaders at IBM anticipate adoption of AI in the corporate world to explode up to 90% in the next 18-24 months.[27] A March 2020 study by Morning Consult and IBM of over 4,500 senior business decision-makers at global companies, finds that across industries and regions, nearly 3 in 4 global businesses have deployed or are ramping up exploratory plans with AI. It finds that global companies – particularly large companies – are planning to invest heavily in all areas of AI over the next 12 months (see Figure 2).[28] While this development and use of AI is not necessarily ML, ML is a common form of AI that has and continues to experience a rapid rise.[29]

For businesses working to deploy AI at scale it is necessary to trust the technology. The aforementioned Morning Consult and IBM study finds that 78% of respondents globally say it is very or critically important that they can trust that their AI's output is fair, safe and reliable.[30] Addressing bias is a critical precursor to establishing this trust.

AI systems are automating judgments

AI models make it possible to automate judgments that were previously made by individuals or teams of people. Using technical systems such as machine learning, AI systems make predictions from data about people and objects related to them.

AI systems are used across the public and private sector. For example, AI systems are used in policing to predict recidivism; in healthcare, education and finance (particularly lending and credit) to serve as diagnostics and/or decision-making aids; and for highly personalized digital advertisements. AI has also become prevalent within business operations. Human Resources (HR) in almost every industry uses AI throughout the hiring process—from determining what audience sees job postings to screening resumes to using video interviews and skill-based online games in screening applicants. AI is used to forecast growth and salaries and inform retention and promotion strategies. AI is impacting the life path of youth too such as through schools using AI to assess student aptitudes and match students to careers or universities.

AI can reduce humans' subjectivity, but AI systems can also embed human and societal biases and produce discriminatory outcomes at scale. The notion that algorithms and AI are unbiased is both inaccurate and indicative of "automation bias", or the over-reliance and over-acceptance for suggestions from systems that are automated.[31] Bias in algorithms and computer systems goes back decades: In 1988, the UK Commission for Racial Equity found a British medical school guilty of discrimination. The school used a computer program that determined which applicants would get interviews. Although it had similar matches to human admission decisions, it was found to be biased against women and those with non-European names.[32]

This playbook is focused on understanding and mitigating bias in those AI systems, which result in inaccurate and/or discriminatory predictions and outputs for certain subsets of the population. The playbook does not focus on other ways biases can manifest in AI systems, such as through the development of gendered voice assistants.[33]

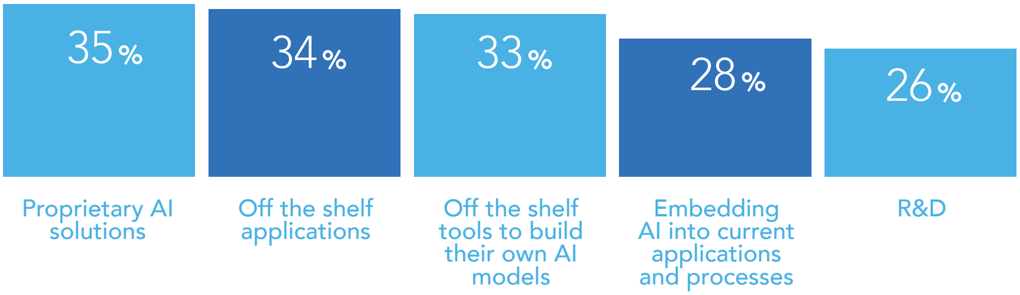


Figure 2. Companies' planned investments in AI over the next 12 months

BOX 1. WHAT IS "BIAS"?

Biases are cognitive shortcuts that can result in judgments which lead to discriminatory practices. While there are various subgroups and types, bias is commonly defined as a "tendency, inclination, or prejudice toward or against something or someone."[34]

Humans experience biases all the time. Our brains are wired to be biased. We have two modes or 'systems' of thinking: System 1 refers to automatic, quick thinking that operates with little to no voluntary control. This system generates our impressions and intuitions, and informs our 'gut' instincts. System 2 involves more deliberate effort and is linked to agency and choice.[35]

We need System 1 thinking to help us organize and manage all the stimuli we constantly face, but this is also where cognitive biases come into play. System 1 thinking relies on associations and categories to discern patterns and make judgments quickly and efficiently. We learn to make associations and categorize based on our personal experiences, education, upbringing and communities -- and the stereotypes and norms that accompany them. Dr. Jennifer Eberhardt sums bias up well: *"Our beliefs and attitudes can become so strongly associated with a category that they are automatically triggered, affecting our behavior and decision making… These associations can take hold of us no matter our values, no matter our conscious beliefs, no matter what kind of person we wish to be in the world."*[36]

AI systems, not beholden to this System 1 thinking, can make predictions and decisions that are less biased and even help identify human biases. While exciting, AI systems are human creations and can still have bias in them (whether the AI system learned this bias from the data it was fed or how it was built). As a result, AI systems can perpetuate and amplify discrimination or marginalization of certain groups – even when they are working as intended.

Often, when referring to "bias" in AI and ML, the term is defined and used in a narrow, technical sense. "Statistical bias" or "algorithm bias" affects the accuracy and reliability of an AI model's prediction. These forms of bias are operationalized technically and require technical fixes. For example, they can come from biased distribution of error rates on the basis of a single variable. These technical forms of biases and technical fixes are important, but don't address all the ways AI systems can be biased. **We define "biased AI" as AI systems that result in inaccurate and/or discriminatory predictions and outputs for certain subsets of the population.**

The concept of bias in AI is tied to fairness. ML systems - particularly those used in policing and crime - can be subject to hot debates about what is "fair". But what's fair can mean different things in different contexts to different people. There are various definitions of fairness across disciplines spanning legal, philosophy, social science and more. The definition of fairness used and the fairness approach taken can inform how bias both manifests and is interpreted. (See EGAL's Fairness Brief for more on what "fairness" means for ML systems, as well as tools and considerations for those developing, managing and using ML systems.

Understand the issue and its impacts

A. Why & how are AI systems biased?

At a high-level, AI systems are biased because they are human creations.They are classification technologies that tend to reflect the dominant culture.[37] Unsurprisingly, they often mirror society.[38]

It matters where AI systems are developed and by whom. The perspectives and knowledge of those who develop AI systems are integrated into them, while the values and priorities of managers and business leaders impact an organization and the products it develops.

Tech companies and labs developing many large scale AI systems tend to be mainly white and male. As shown in Figure 3, in 2013, the share of women in computing dropped to 26%, below the level

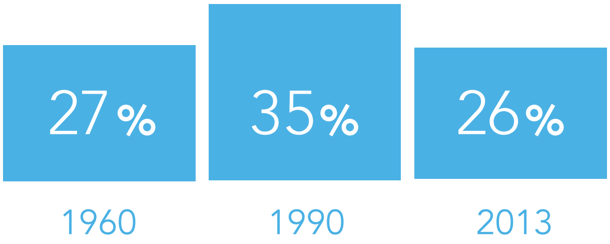


Figure 3. Percent of women in computing decrease below 1960 threshold

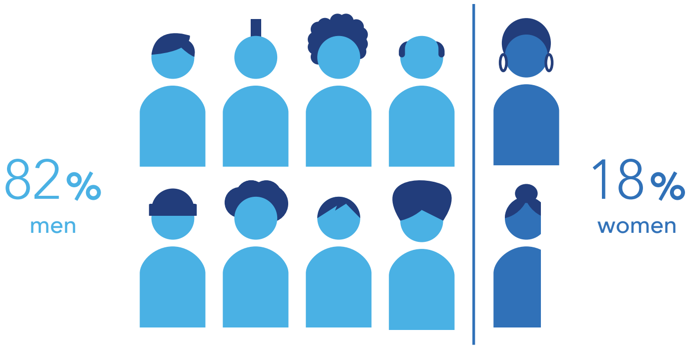


Figure 2. **Only 18% of researchers at leading AI conferences are women**

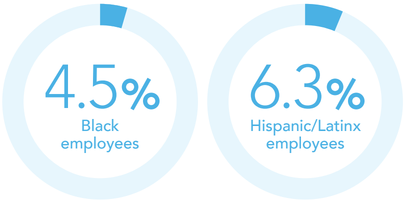


Figure 3. **At a leading US tech company only 4.5% of employees are Black, and only 6.3% are Hispanic / Latinx**

it was at in 1960.[39] Further, almost half of women who go into tech eventually leave the field, which is more than double the percentage of men who leave.[40] In AI research, only 18% of authors at the leading 21 conferences are women[41] and 80% of AI professors are men (see Figure 4).[42] This imbalance is replicated at large technology firms. There is lack of reported data on transgender or other gender minorities.[43] Racial diversity is also lacking. As seen in Figure 5, at Microsoft, 4.5% of employees are Black and 6.3% are Hispanic/Latinx.[44] These numbers are similar at other technology firms.

Beyond demographic diversity, in many cases AI systems are not designed with relevant domain experts nor informed by end users.

More specifically… bias can be present in the generation, collection and labeling / management of data that the algorithm learns from, as well as the design and operation of algorithms. Our Bias in AI Map breaks this down.

The Bias in AI Map

Our **Bias in AI Map** breaks down where and how bias can enter in datasets and algorithms if action and intervention is not taken. Figure 6 highlights the high-level view of the Map. We then zoom into the **pathways to a biased dataset** (visual 1) and the **pathways to a biased algorithm** (visual 2). We also illustrate how **use of an AI system** matters (visual 3). Examples are provided across different industries and use cases throughout.

Data incorporates societal inequities and cultural prejudices. Vast amounts of data points are **generated** by virtue of individuals' day-to-day activities (e.g., consumer behavior, health conditions) and data points are **collected** through various platforms, technological or otherwise. Data is assumed to accurately reflect the world, but there are significant data gaps (including little or no data coming from particular communities[45]) and data is rife with racial, economic and gender biases.[46] In addition, human influence cannot be eliminated from data. In many cases, humans decide what, where and how data is collected and categorized,

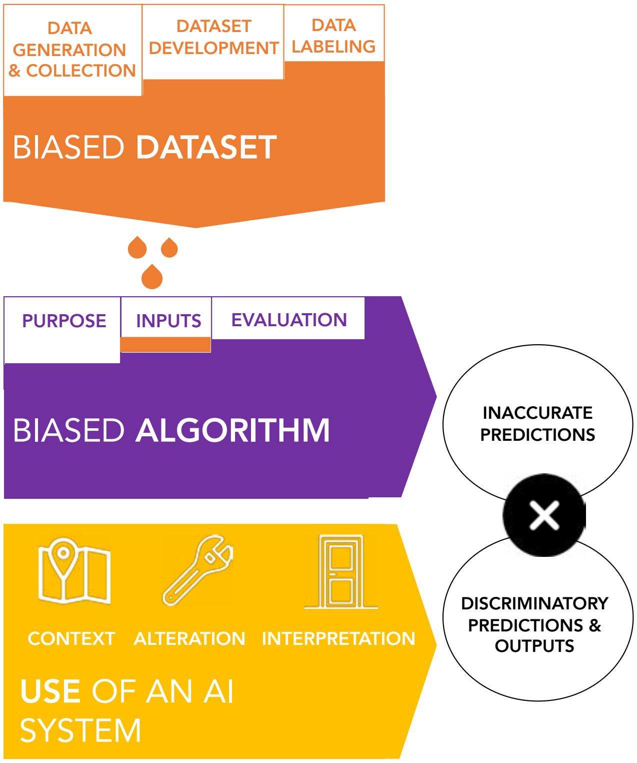
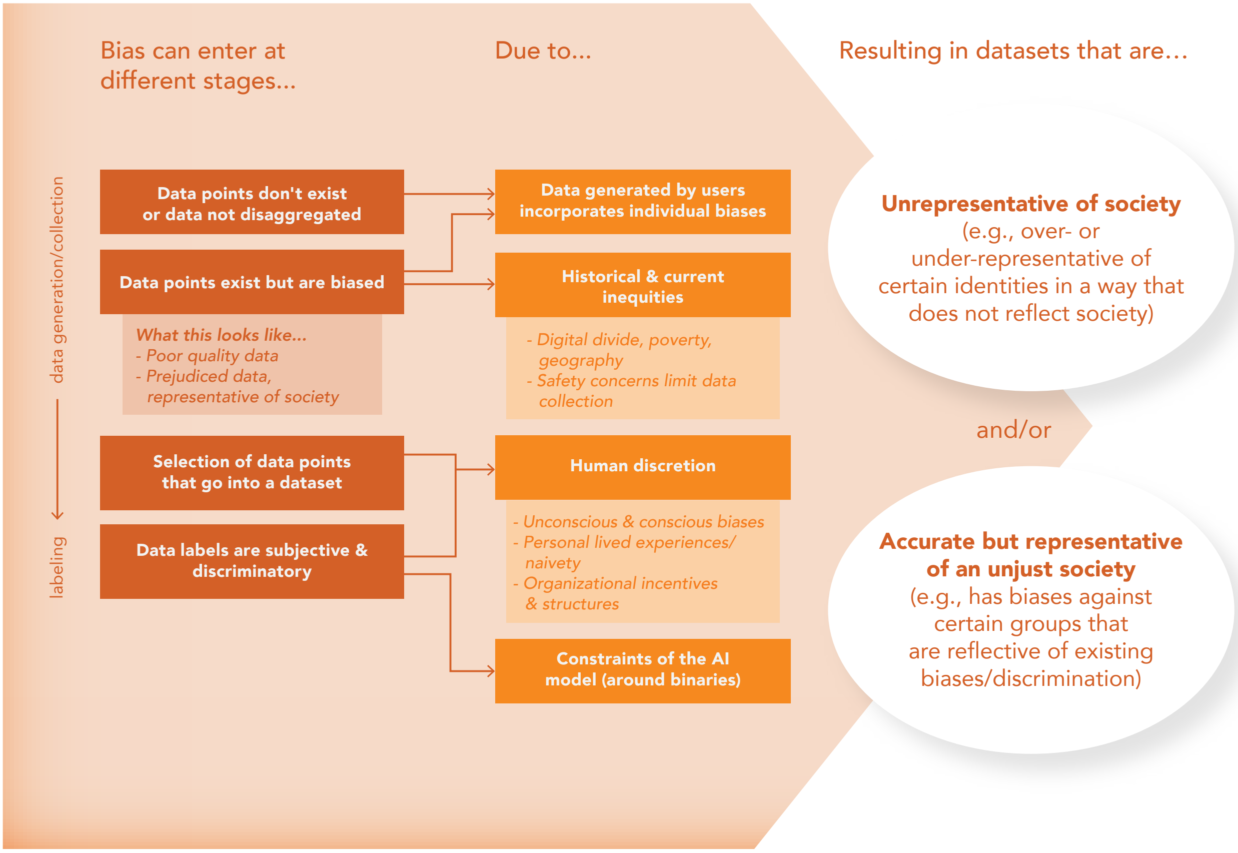


Figure 6. Bias in AI Map from a high-level view shows how AI systems can result in inaccurate or discriminatory predictions and outputs.

VISUAL 1. PATHWAYS TO A BIASED DATASET*



Bias can enter at different stages...

Due to...

Resulting in datasets that are...

data generation/collection → labeling

Data points don't exist or data not disaggregated

Data points exist but are biased

What this looks like...
- Poor quality data
- Prejudiced data, representative of society

Selection of data points that go into a dataset

Data labels are subjective & discriminatory

Data generated by users incorporates individual biases

Historical & current inequities

- Digital divide, poverty, geography
- Safety concerns limit data collection

Human discretion

- Unconscious & conscious biases
- Personal lived experiences/ naivety
- Organizational incentives & structures

Constraints of the AI model (around binaries)

Unrepresentative of society (e.g., over- or under-representative of certain identities in a way that does not reflect society)

and/or

Accurate but representative of an unjust society (e.g., has biases against certain groups that are reflective of existing biases/discrimination) as well as parameters for a **dataset**. Data is also **labeled**, which can be subjective. Much data was collected at some time in the past, and datasets then preserve that past.[47]

Datasets can then be biased by over- or under-representing certain identities in a particular context that does not reflect reality. Biased datasets can also be accurate but representative of an unjust society. In this case, they reflect biases against certain groups that are reflective of real discrimination that the particular group(s) face.

Understanding bias – and addressing it – in data requires a thorough understanding of the social, political and economic context and conditions through which the data was produced. It also requires a sensitivity to the methods and instruments of data generation and collection. Data is not objective – claims that it is objective ignores the lived history of data and poses a barrier to mitigating it.

Let's dig into why and how a dataset can be biased, looking closer at each of the steps data has to go through before being ready for an algorithm to learn from.

> **"Data is never this raw, truthful input and never neutral. It is information that has been collected in certain ways by certain actors and institutions for certain reasons."**
>
> *Catherine D'Ignazio, Assistant Professor at Massachusetts Institute of Technology (MIT)* [53]

1. Data generation and collection

In some cases, **data points may not exist** for certain groups, identities or communities. This can be due to *historical and current inequities.* This includes certain groups or individuals having less access to technological tools that would generate / collect data.[48] Technologies are always differentially adopted and any divide in accessing digital technology is not a one-time event, but a constantly moving target as new devices, software and cultural practices emerge. If data is generated by people on the Internet or through other technologies, then it will inherently not be representative of the whole world and different groups.[49] Technologies may have been designed in ways that exclude certain populations purposefully or not. People may also live on big data's margins due to poverty, geography or lifestyle.[50] In some cases, individuals or communities will resist data being collected on them out of safety concerns and/or for fear of repercussions and exploitation.

Some 2.3 billion women worldwide do not have any Internet access, and more than 1.7 billion do not own a mobile phone—some 200 million fewer women than men have online access or mobile phones.[51]

As in many communities of color hard-hit by Covid-19, immigrants are at higher risk for exposure to the virus as they are filling "essential" positions. Many immigrants are also at increased risk of complications or death from COVID-19 due to high rates of underlying chronic illnesses such as heart and lung disease. But many are also not getting tested for fear of being deported.[52]

BOX 2. A HISTORY SNAPSHOT OF OUR DATAFIED WORLD

The history of data is bound up with the history of European colonialism and capitalism. Data has been emerging largely since the 17th century alongside European colonialism and expansion. As states were consolidating power over controlled territories they increasingly obtained and classified information about those territories. Data collected by these European powers became information that then became history. This data and history are tied to the creation of race and racism, and is also deeply gendered.[60]

Furthermore, *individuals or groups that are collecting data make choices* on what data to collect and how. Irresponsible questions and data collection methods can lead to lack of data being collected for certain groups, particularly those that have been traditionally marginalized. Who is collecting the data matters – lived experiences inform decisions, perspectives and what might be missed or overlooked.

While nearly 40% of Americans identify as being nonwhite, 80-90% of participants in most clinical trials are White.[57] Further, in 2015, only 1.9% of respiratory disease studies included any minorities.[58]

As of 2016, >80% of all genetics data is from individuals of European ancestry.[59]

Healthcare is rife with examples: Men's bodies have always been the standard for medical testing. Women are missing from medical trials with female bodies deemed too complex and too variable. Researchers also note that women tend to be harder to schedule for tests in studies, likely given caretaking responsibilities. Females aren't even included in animal studies on female prevalent diseases. Some researchers continue to advocate against including women in research on the basis that while biological sex may matter, lack of comparable data arising from historical data gaps.[54] The result? FDA approved drugs[55] with greater health risks for women and more misdiagnoses and fatalities of women from heart attacks.[56]

Data not disaggregated by sex, gender, race, ethnicity, etc. may paint an inaccurate picture for particular identities. It can conceal important differences between subgroups, and hide potential over- or under-representation of certain groups.

Few urban datasets track and trend data on gender so it is hard to develop infrastructure programs that factor in women's needs. Urban planning has failed to account for women's risks of being sexually assaulted, reflected in the fact that 33% of women globally lack access to safe toilets and women are twice as likely as men to be scared in public places. Fear of crime is particularly high among low income women and ethnic minority women.[61]

It was only in mid-April that an update released by the CDC contained a race and sex breakdown of data on COVID-19 cases and deaths – and it only pulled from hospital networks in parts of 14 states.[62] This is important because men and women are likely to have fundamentally different reactions to the virus, vaccines and treatment.[63]

When data points do exist, they may have prejudice built in and reflect inequities in society. Although data may be accurate and reflect genuine differences (e.g., men tend to be taller than women) some biases may masquerade as 'genuine differences' but be rooted in societal bias.[64] In many cases data is not objective, but reflects pre-existing bias linked to *historical and current inequities*. Algorithms using this data can create dangerous feedback loops.

This can also come from *data generated by users* creating feedback loops that lead to bias.[65]

Black Americans are more likely to be arrested and incarcerated in the U.S. due to historical racism, disparities in policing practices, or other inequalities within the criminal justice system.[66] Historic and current higher arrest and incarceration rates are fed into AI systems that then perpetuate them.

Natural Language Processing algorithms—rapidly gaining popularity in a variety of fields—rely on previously published text corpora that are available online. Research on word embeddings over the last 100 years shows that biased language around gender and ethnic stereotypes are replicated when the tool is asked to make new associations.[67] Occupations such as "nurse" and "homemaker" are more likely to be assigned a female gender than occupations such as "doctor", "computer programmer". These biased results then become new inputs that create a feedback loop which doesn't consider changing norms and trends in present society.[68]

In Latanya Sweeney's research on racial differences in online ad targeting, searches for African-American-identifying names tended to result in more ads featuring the word "arrest" than searches for White identifying names. Sweeney hypothesized that even if different versions of the ad copy – versions with and without "arrest" – were initially displayed equally, users may have clicked on different versions more frequently for different searches, leading the algorithm to display them more often.[69]

It can also come from the *involvement of human discretion* in defining what data to generate / collect. Irresponsible data collection methods that lack context, contain critical omissions and misdirected questions can lead to poor quality data. Such methodology and reporting errors can compound the exclusion experienced by marginalized communities.

Federal and state authorities have struggled to capture labor force data from Native American communities, and data has been riddled with methodology errors and reporting problems. Data on Native Americans is often not standardized and different government databases identify tribal members at least seven different ways using different criteria – with federal and state statistics often misclassifying race and ethnicity as well. This has resulted in undercounting of the American Indian population and key indicators like joblessness.[73]

2. Selection of data points to go into a dataset

Irresponsible data selection for the development of a dataset is linked to *involvement of human discretion* in the selection of data points that go into a dataset. These individuals may make choices that are naive of the societal context, power structures and historical inequities that can impact the dataset, or may (inadvertently or not) enter their own biases. Who is collecting the data matters – lived experiences can inform decisions, perspectives and what might be missed or overlooked. Datasets may therefore be of 'poor quality', in that they over- or under-represent a certain group or population or inaccurately convey a certain group, identity or population. Or more simply, the individual or team does not prioritize or ensure representation in the dataset.

Research from Joy Buolamwini and Timnit Gebru highlighted the lack of diverse and representative samples in image datasets used for commercial facial recognition systems, that then misclassified gender far more often when presented with darker skinned women compared with lighter skinned men (error rate of 35% versus .8%).[70] Following the results of this research, companies responded to address the discrepancies. For example, IBM increased the accuracy of its facial analysis by using broader training datasets and more robust recognition capabilities to achieve a nearly ten-fold decrease in error-rate.[71]

A dataset used to train neural networks to identify skin cancer from photographs had <5% images of dark-skinned individuals.[72]

3. Data Labeling

Labeling of data *involves human discretion* and can be subjective and discriminatory building from harmful biases, naivety, priorities and perspectives. Labeling of images can be particularly problematic. Collecting, categorizing and labeling images "is itself a form of politics, filled with questions about who gets to decide what images mean and what kind of social and political work those representations perform."[74]

When it comes to assigning gender labels, most data end up classified in terms of simplistic, binary female / male categories. When gender classification collapses gender in this way, it reduces the potential for gender fluidity and self-held gender identity.[75] As a result, a simplistic cultural view of gender is built into tools such as image classifications and facial analysis systems, leading to the erasure of nuanced identities that are already marginalized by society. Disability – which encompasses a number of physical and mental conditions each with its own history and specificity – also resists fitting into neat arrangements and fixed classifications. The boundaries of disability continually shift and can't be added as one more stand-alone axis of analysis.[76] Challenges around this as well as potential solutions are important to explore further to ensure AI systems aren't reiterating erasure of certain identities.[77]

A Beauty AI competition trained robots on pre-labeled images of what was 'beautiful'. The robots did not like people with dark skin.[78]

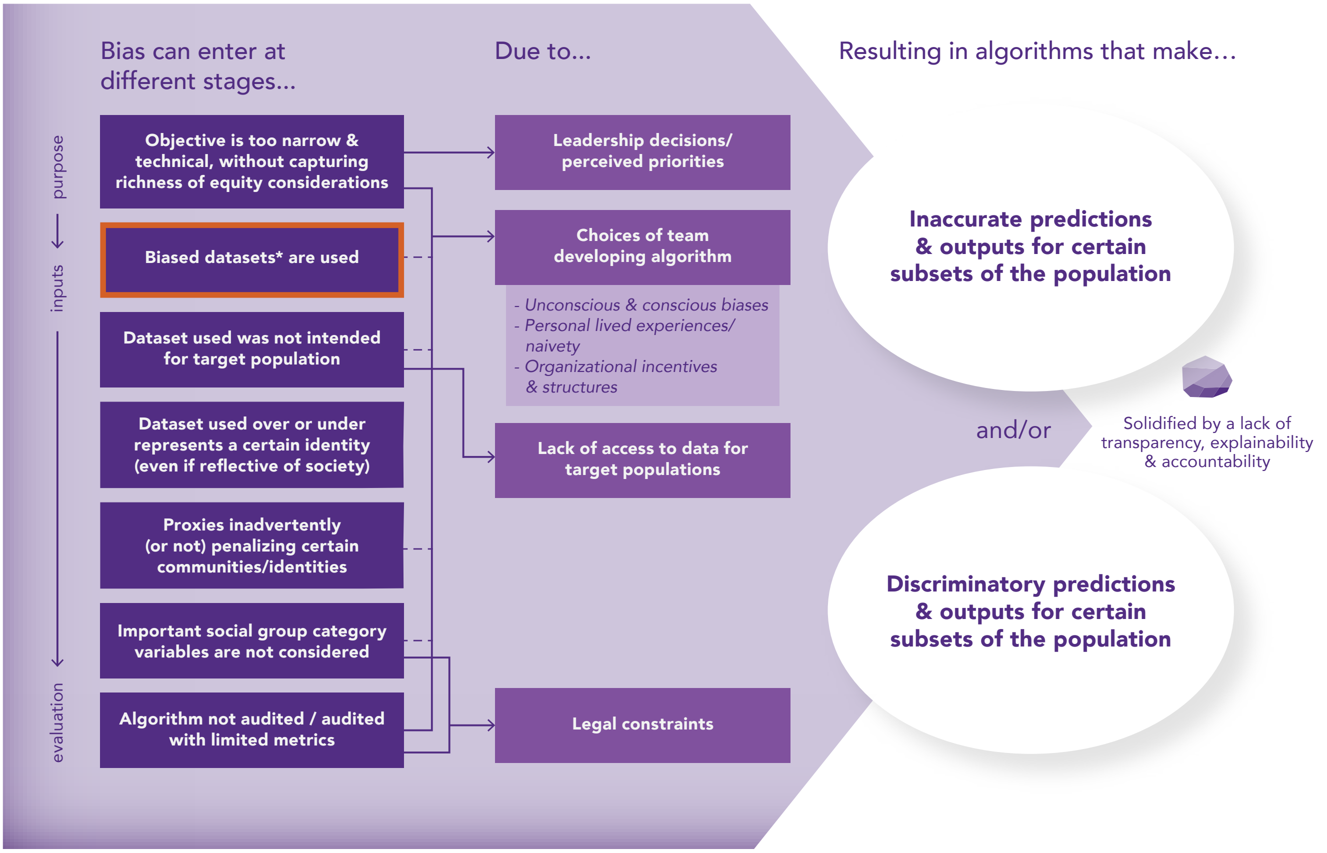
Photo: The Gender Spectrum Collection

Although algorithms make decisions based on the data they are fed, human input is required to define the **purpose** of an AI model and the constraints it operates under, as well as the **inputs** it should consider in order to find patterns and draw conclusions. An algorithm can lead to technically inaccurate predictions for certain subsets of the population, particularly if drawing from a dataset in which a certain identity is over- or under-represented in the dataset or if a dataset is used that is not specific to the algorithm's target population. An algorithm can contribute towards discriminatory outputs, irrespective of the quality of data and training dataset that is used, and how algorithms are **evaluated** is important. Designers and operators of algorithms should watch for potential negative feedback loops that can cause an algorithm to become increasingly biased over time.

Humans ultimately have their own deeply entrenched biases, and may think what they are doing is neutral and scientific but actually is perpetuating or enhancing prejudice and inequities.

Let's dig into why and how algorithms can be biased, beginning with the purpose of the algorithm through the algorithm's inputs.

VISUAL 2. PATHWAYS TO A BIASED ALGORITHM


Bias can enter at different stages...
purpose
inputs
evaluation
Objective is too narrow & technical, without capturing richness of equity considerations
Biased datasets* are used
Dataset used was not intended for target population
Dataset used over or under represents a certain identity (even if reflective of society)
Proxies inadvertently (or not) penalizing certain communities/identities
Important social group category variables are not considered
Algorithm not audited / audited with limited metrics
Due to...
Leadership decisions/ perceived priorities
Choices of team developing algorithm
- Unconscious & conscious biases
- Personal lived experiences/ naivety
- Organizational incentives & structures
Lack of access to data for target populations
Legal constraints
Resulting in algorithms that make...
Inaccurate predictions & outputs for certain subsets of the population
and/or
Solidified by a lack of transparency, explainability & accountability
Discriminatory predictions & outputs for certain subsets of the population

1. Algorithm purpose

The purpose of the algorithm is key. Most AI systems are trained to optimize for specific objectives that are fairly narrow to maximize accuracy for a particular prediction path.[79] This narrow purpose can lead to a 'value alignment problem' between the problems the AI system begins to solve and human values.[80] Running an ML algorithm with the sole objective of minimizing error rates (as is done typically), can mean missing other mistakes like how error rates differ for different demographics. The purpose is informed by *leadership decisions and perceived priorities*. It can also be linked to the *team of algorithm developers* and their personal biases or lack of understanding the broader context in focus.

A group of researchers explored how college admission algorithms that make decisions by predicting GPA of admitted students might be improved with equity as the priority versus efficiency. When the authors use an "equity lens" (that considers race), as opposed to an "efficiency lens" (that only considers predicted grades), the algorithm does better in admitting a more equitable percentage of White and Black students and it does better in leading to admitted students with higher grades on average.[81]

The COMPAS algorithm developed by the firm Equivant (formerly Northpointe) forecasts which criminals are most likely to reoffend. It was designed with the purpose of correctly predicting recidivism for defendants and being as accurate as possible across all individuals. While it did correctly predict recidivism for Black and White defendants at roughly the same rate, when it was wrong, it was wrong in different ways for Black and White people: Black arrestees who would not be rearrested in a 2-year horizon scored as high risk at twice the rate of White arrestees not subsequently arrested. It also scored White people who were more likely than Black people to go on to commit a crime, as lower.[82] By doing this, the algorithm perpetuates a status quo, without incorporating how and why the policing system has and continues to be discriminatory against Black people.

2. Algorithm inputs

2a. Dataset use:

Automated processes require agreement on what data is relevant to a given decision. The selection of the dataset(s) to train the algorithm and help solve the problem the AI system will be tackling is critical. Bias can enter datasets in a variety of ways, as we outlined previously (see visual 1 and pathways to a biased dataset). By selecting and using a *biased dataset,* the team developing the algorithm builds in those biases. A team could also select and use *a dataset out of context and/or not specific to the algorithm's target population*. This could be due to irresponsibility of those selecting the datasets or lack of access to data / a dataset for the target population(s). Lastly, algorithm developers could select a *dataset that over or under-represents a certain identity* (even if it's reflective of society). In this case, the dataset may have less data on a minority group in a particular community or context. At the same time, collecting more data on a minority group can also link to issues of ethics and privacy.

In 2018, Amazon decided to recall its experimental AI-driven hiring tool which was designed to mechanize the search for top talent when it was found to be biased against women, penalizing candidates whose resumes included the word "women's".[83] The model was trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period—most of whom were men. The tool drew from a dataset that reflected historical inequities and gender disparities in the tech sector. The preference for male candidates was carried over in evaluating new candidates, in effect exacerbating the bias instead of eliminating it.

2b. Proxies / variables:

Proxies[84] selected for the algorithm may penalize certain identities or communities. While this often happens inadvertently, it can also be purposeful and be tied to financial interests (see example on affinity profiling to the right). This links back to the decision of the teams(s) developing the algorithms and the data used – which may be linked to and/or reflective of historical inequities and prejudices to begin with.

Affinity profiling (the practice of grouping people based on their assumed interests rather than on their personal traits) is common in online advertising, but has potential to be discriminatory if people do not see certain ads or receive different prices based on their affinity. As recently as 2016, Facebook allowed advertisers to use "ethnic affinity" as a proxy by which to target people by race. In 2018, Facebook removed 5,000 affinity categories to address concern from activists and lawmakers, while also adding requirements that advertisers behind ads for housing, jobs and credit cards comply with a non-discrimination policy.[85]

An online tech hiring platform, Gild, enables employers to use 'social data' (in addition to other resources such as resumes) to rank candidates by social capital. Essentially 'social data' is a proxy that refers to how integral a programmer is to the digital community drawing from time spent sharing and developing code on development platforms such as GitHub. This selection ignores key societal context: societal expectations around unpaid care put greater time burdens on women so they have less time to chat online. It also ignores how women may assume male identities given sexist tones on platforms like GitHub, gender-specific safety concerns (e.g., targeted harassment and trolling), and other forms of bias. For example, while women's contributions tend to be accepted more often than men's on GitHub, women's acceptance rates are higher only when they don't identify as women.[86] Instead of removing human biases, Gild ultimately created an algorithm with hidden gender bias.[87]

In one algorithm widely used in the US, researchers found it falsely concluded that Black patients are healthier than equally sick White patients. This reduced the number of Black patients identified for extra care by more than half. The algorithm uses health costs as a proxy for health needs, but Black patients who have the same level of need spend less money because of unequal access to care, among other reasons. Reformulating the algorithm to no longer use costs as a proxy for needs more than doubles the number of Black patients receiving adequate care.[88]

On the other hand, algorithms that may not incorporate explicit proxies or variables of certain *social group categories* (e.g., race, ethnicity, sex, gender) can still discriminate based on these categories. Algorithms may not incorporate this information because – depending on the industry and country context – it may be illegal to incorporate social group categories as they are legally protected classes. However, without social group category data, one can ignore or hide – rather than prevent – discrimination. Even if not explicitly using a social group category, an algorithm may pick up on statistical correlations that are socially unacceptable or illegal. Also, social category data are often needed to check whether discrimination is taking place.[89] In many cases, integrating social group categories is important for more accurate and equitable outputs.[90] Not incorporating race and gender explicitly simply masks unequal histories of market exclusion, devaluation of labor, and other inequities. Worse, the history of discriminatory practices is harder to see and confront in the automated system as it's seen as "objective" and "fair".[91]

In the consumer credit industry, early processes used variables such as clothing, gender, and race to determine creditworthiness. Eventually these discriminatory variables were replaced by ones considered more neutral and directly relevant to judging likelihood to repay (e.g., information about financial history). But by then, women and people of color had less formal financial history and suffered from discrimination, impacting their ability to get credit. AI systems that determine creditworthiness reproduce the same inequitable access to credit along race and gender lines.

Research by Berkeley Haas professor, Adair Morse, on bias in the algorithmic scoring used by FinTech lenders in this market finds that it results in minorities paying $765 million more interest annually on existing stocks of mortgage.

A research study found that keeping race in admission decisions for colleges improves predicted GPA of admitted students and increased fraction of admitted students who are Black. On the other hand, race-blind predictors mis-rank Black students. A reason for this might be that Black students do not have the same access to resources such as SAT preparation.[92]

3. Algorithm evaluation

Auditing is an important part of the process for evaluating and iterating an algorithm that informs if the code has issues and where. *Not auditing or auditing only for certain metrics invites unwanted and harmful bias.* The metrics that one uses for defining success and auditing are important. If the team is only looking at accuracy and not false positives across different demographics, then the audit can hide a lot of error types. Harm can be introduced by not testing certain conditions.

Criminal justice risk assessment tools are often created outside of the jurisdictions where they are deployed. These types of "off-the-shelf" products are not developed collaboratively with the communities they impact, or tailored for the conditions and contexts in which they are used.[93]

BOX 3. AI AS A TOOL TO EXPLICITLY MITIGATE BIASES[102]

As AI systems make it possible to automate judgments that were previously made by individuals or teams of people, they can reduce error rates associated with decision making by identifying and/or mitigating humans' own biases[94]. There are various examples from different industries and sectors where AI is being used to explicitly tackle biases.

In lending, UC Berkeley professor Paul Gertler with Sean Higgins (Northwestern University) and Laura Chioda (World Bank) are exploring how algorithms can mitigate gender bias for low-income communities. Although women often have better repayment rates for loans, they often struggle to access traditional loans because of gendered discrimination linked to not legally owning assets for collateral, not having credit / earnings histories and/or discrimination from loan officers.[95] Working with a bank in the Dominican Republic, the researchers are using gender-differentiated credit lending algorithms that make credit decisions for men and women using mobile bank account information. Initial results show that this gender-differentiated approach is promising and also highlights the importance of being able to incorporate social group categories such as gender for more equitable outcomes.[96]

In governance and justice, TrialWatch[97] (a program launched in May 2019 by The Clooney Foundation for Justice and Microsoft's AI for Good initiative) monitors trials globally to highlight injustices and rally support for defendants whose rights have been violated in the criminal justice process. The program recruits and trains people to monitor trials, and then has these individuals use an AI-enabled app to capture audio (speech-to-text translation and language translation), photos and questionnaires. The AI makes information comparable and teases out trends. Leveraging this data and information, TrialWatch works with human rights experts to assess fairness of trials and shares reports and dossiers with other stakeholders.

In healthcare, IBM is using computer vision technology with AI to eliminate skin cancer misdiagnoses that disproportionately impact dark-skinned people. While light-skinned Americans have a higher risk of developing skin cancer than African Americans (who are 22 times less likely), survival rates for African-Americans are much lower (77% vs. 91% for Caucasians).[98] The AI computer vision tool is trained on a dataset of various skin types and learns from historical diagnoses to identify areas that look like lesions for further analysis and make diagnoses. Currently, this tool is equally as accurate as a specialist at recognizing melanoma across a visual dataset.[99]

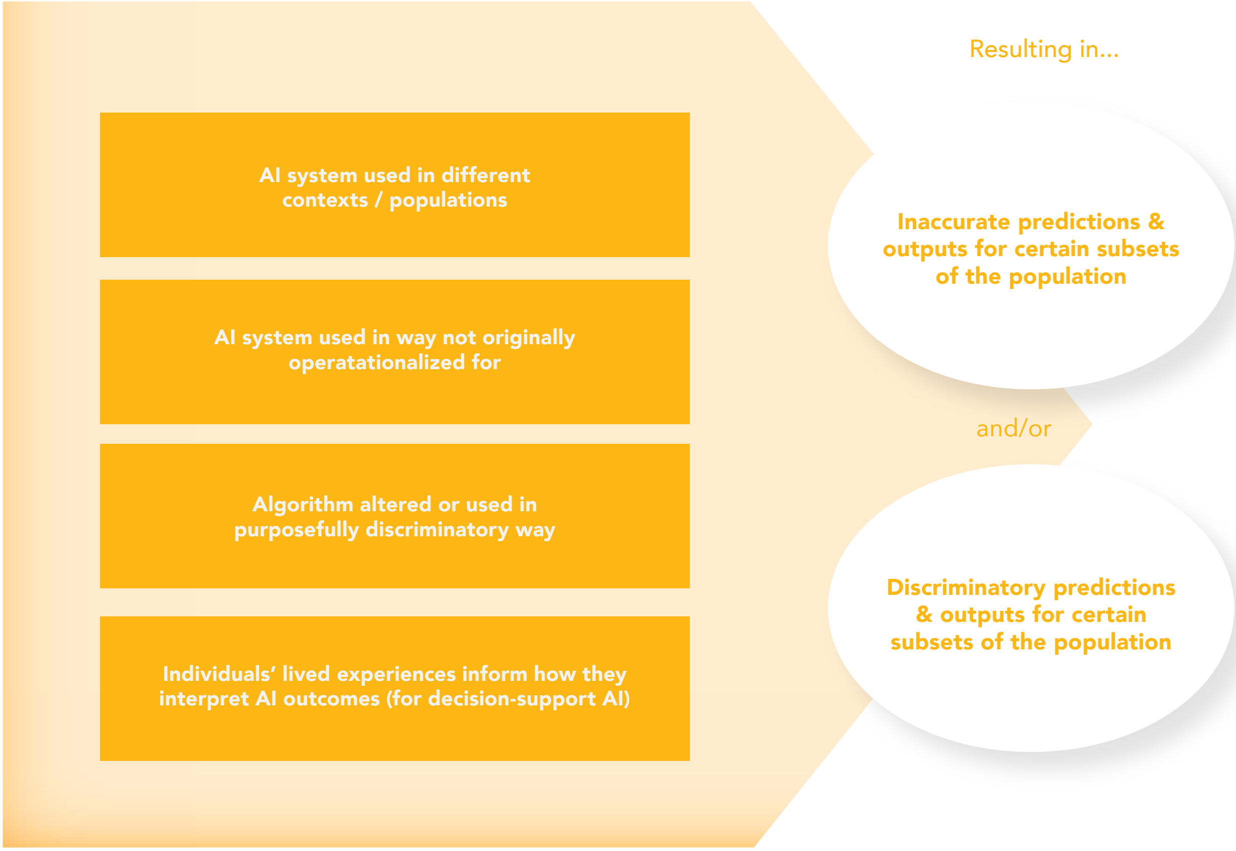
In hiring, there is an influx of AI-driven recruitment tools and startups that seek to tackle "similarity attraction" effects and confirmation biases that exist in traditional recruiting and interview processes. AI is designed to select candidates objectively by making screening decisions based on relevant data points. Research on performance of a job-screening algorithm at a software company finds that it did favor "nontraditional" candidates at a higher rate than human screeners did, exhibiting significantly less bias (although not free of bias) against candidates that were underrepresented at the firm.[100]

While promising, technology products and services that offer to fix societal bias can still reproduce or deepen discriminatory processes. These types of AI systems can still be exposed to certain pathways for bias to enter – in particular, they can be at risk for using datasets that are accurate but representative of an unjust society. It is critical to pay close attention to this risk and assess the various potential ways these systems could inadvertently do this prior to the development of an AI system. Using the **Bias in AI Map** as a guide can help. It can be challenging to know where and how social prejudices have been built into a technology. As author Ruha Benjamin notes: *"The practice of codifying existing social prejudices into a technical system is even harder to detect when the stated purpose of a particular technology is to override human prejudice."*[101]

Beyond a biased algorithm(s), AI systems can result in discriminatory outcomes for certain individuals or populations based on how they are used. There is potential for inaccurate predictions and bias if an AI system is *used in a different context or for a different population* from which it was originally developed or if it is *applied for different use cases from which it was originally developed / operationalized.*[103] This can be problematic if the original AI system doesn't capture changing societal knowledge (expertise, habits) or population values.

AI systems can be *used or altered by organizations or individuals in ways that can be deemed as discriminatory* for certain populations. This can be due to bad actors getting ahold of and using the technology. In other cases, it may be less overt and subject to debates over fairness. Lastly, for AI systems that support human decision making, how *individuals interpret the machine's outputs* can be informed by one's own lived experience and allow for bias.

VISUAL 3. USE OF AI SYSTEMS


AI system used in different contexts / populations
AI system used in way not originally operatationalized for
Algorithm altered or used in purposefully discriminatory way
Individuals' lived experiences inform how they interpret AI outcomes (for decision-support AI)
Resulting in...
Inaccurate predictions & outputs for certain subsets of the population
and/or
Discriminatory predictions & outputs for certain subsets of the population

Clearview AI has a massive database of 3 billion photos scraped from social media without permission or knowledge of individuals. The company maintains that its facial recognition tool was meant to be used only by law enforcement and a few private companies. But the company is not transparent or upfront about how the technology is being used and by whom.[104] It is also being used by the U.S. government and other law enforcement agencies without proper knowledge about how the technology works and who is behind it. There are immense "weaponization possibilities" of using such secretive AI.[105] The threats to privacy posed by Clearview AI are becoming evident – in May 2020, the American Civil Liberties Union (ACLU) sued the company on behalf of vulnerable communities that are likely to be harmed by its surveillance capabilities. New York filed a separate class action against the firm in January.[106]

In 2013, the US Immigration and Customs Enforcement (ICE) started using a software tool to recommend whether people arrested over immigrant violations should be let go. The New York Civil Liberties Union (NYCLU) and Bronx Defenders allege that the algorithm is unconstitutional as it detains most individuals even if deemed a minimal threat to public safety. The NYCLU found that the algorithm had been altered to increase detention without bond of "low risk" individuals. The risk assessment tool was also altered in 2015 to remove the possibility of a "release" output and remove the option for bond.[107]

Predictive policing tools are often used off-the-shelf without being adjusted for the particular context they are used in, opening up potential for inaccurate prediction and bias in the new context. For risk assessment tools used by police and judges, it's important to understand how judges, police officers and other decision makers interpret its results.

B. What are the impacts of biased AI?

...For individuals and society

Biased AI systems have immense impacts on the lives of individuals. AI systems can "unfairly" allocate opportunities, resources or information. This may worsen inequities experienced by individuals in underserved communities. From an economic point of view, discriminatory resource allocation leads to inefficiencies and losses in the economy / markets.

Biased AI systems can pose a detriment to the safety of individuals. They can also fail to provide the same quality of service to some people as others and negatively impact a person's wellbeing such as by being derogatory or offensive, or treating people as if they don't exist. Not allowing for different gender identities and nuanced understandings of (dis)ability contributes towards treating people as if they don't exist, which also has implications for the wellbeing of individuals. Impacts can be of varying severity – sometimes seemingly non-severe harms can accumulate and be extremely burdensome.[109]

AI systems can infringe on civil liberties, while also reinforcing existing prejudices. Given that decision-making processes of ML algorithms cannot often be fully mapped out or understood by humans (referred to as "black box"), it makes it harder for affected individuals to argue that they've been discriminated against.

At a larger societal level, when unchecked, AI systems can solidify and amplify societal discrimination. Given people's tendencies to favour automated decision-making systems and trust that they are inherently neutral and "objective" (aka automation bias), unchecked biases encoded in algorithms get routed through technoscience and coded as "scientific".[110] Research shows that when exposed to consistently biased algorithmic outcomes, consumers are likely to use these results to confirm / reinforce their own biased perceptions of the world.[111] A study of images in 2017 revealed that pictures of cooking were over 33% more likely to involve women than men. However, algorithms trained on this data connected pictures of kitchens with women 68% of the time (See Figure 7). It was also found that the higher the original bias, the stronger the amplification effect – which explains how an algorithm came to label a photo of a "portly balding man in front of a stove" as female.[112] This feedback and amplification effect is persistent in various concerning areas, such as policing and hiring.

As AI systems used by government officials are often developed by private companies, private companies have the power to act like political entities without the checks and balances. ML systems used by the government are often developed by private companies (e.g., COMPAS for policing was developed by the company

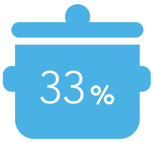
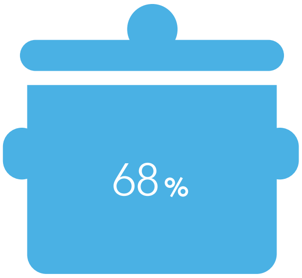


Figure 7. A study of images in 2017 revealed that pictures of cooking were over 33% more likely to involve women than men



However, algorithms trained on this data connected pictures of kitchens with women 68% of the time.

Equivant (formally Northpointe)) so government officials are essentially outsourcing decisions that should be the purview of democratic oversight.

...For business

Biased AI systems can result in reputational costs for the companies that produce and/or use them, with implications for undermining the AI systems, having to do damage control, as well as losing consumers and future market opportunities. Large tech companies recognize this risk: Microsoft flagged reputational harm or liability due to biased AI systems as a risk to their business in a report to the US Securities and Exchange Commission.[113] A 2019 DataRobot report found that 42% of organizations currently using / producing AI systems are "very" to "extremely" concerned about the reputational damage that media coverage of biased AI can cause (see Figure 8), with most respondents citing "compromised brand reputation" and "loss of customer trust" as their greatest cause for concern.[114] For companies that produce AI systems that are bought by individuals or other organizations, this can impact sales. Edelman research from 2019 finds that three-fourths of consumers today won't buy from unethical companies, and 86% say they're more loyal to ethical companies.[115]

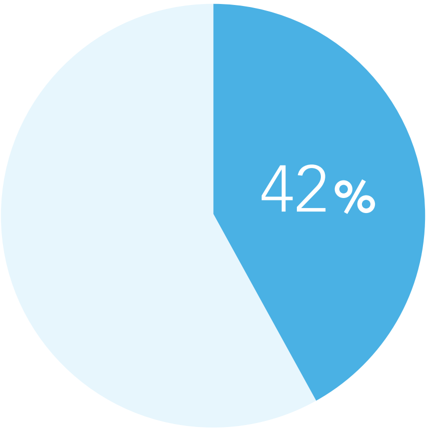


Figure 8. A 2019 DataRobot report found that **42% of organizations currently using / producing AI systems are "very" to "extremely" concerned** about the reputational damage that media coverage of biased AI can cause.[108]

Employee demand for more ethical practices around AI has implications for internal conflicts and unwanted media attention that could damage corporate reputations. On the other hand, ethical companies can better attract and retain talent.[116] In 2018, employees at Google staged protests and walkouts to showcase their opposition to the use of its drone analysis AI by the Pentagon.[117] As a result, the CEO announced that the contract would not be renewed, and that the company would no longer develop AI weapons or technologies that may be weaponized to violate internationally accepted norms. Along a similar vein, employees at Amazon wrote a letter to their CEO expressing their concerns over the biases found by the American Civil Liberties Union in the company's facial recognition algorithm.[118]

AI systems found to be biased may be scrapped or need significant changes, resulting in high costs in terms of employees time and other invested resources. For instance, following immense backlash from employees and policy makers, Amazon placed a one-year moratorium on police use of its facial recognition technology, Rekognition. This came on the heels of similar decisions by other tech giants such as IBM,

which discontinued its general-purpose facial recognition system, and Microsoft, which announced that it would stop selling its system to police departments until federal law regulates the technology.[119] This doesn't just affect large companies – it can result in major losses for start-ups whose main product offering is based on a single premise as well. For instance, in 2014, two tech entrepreneurs launched an app called SketchFactor, which would allow users to report having experienced something "sketchy" in a particular location, and the resulting reports would be geotagged and overlaid on a map to create a 'sketchiness heat map' of sorts. The app was immediately met with backlash for propagating teleological redlining, and establishing a "rating system" that didn't account for the systemic racism prevalent in the country. Eventually, the app had to be scrapped, and the entrepreneurs were forced to pivot entirely.[120]

Biased AI systems that serve as decision support systems can face resistance from professionals that use them as well, limiting their value proposition. These professionals might not feel tools capture professional judgment, or feel comfortable using an opaque system, for example, resulting in protest or resistance to their use.[121]

Lastly, companies that don't make addressing bias in AI a priority may be liable to incur large penalty fees. The extent of these fees vary across different industries, however there are some trends within and across industries that legislation is advancing. The Algorithmic Accountability Act, a bill introduced to the US Senate in 2019, is a clear sign that such regulations are in the pipeline. Related to facial recognition technology, in June 2020 democratic lawmakers introduced a bill that would ban federal agencies – including law enforcement - from using facial recognition technology for surveillance. The proposed bill comes amid heated debate over policing and racial justice, on the heels of Black Lives Matter protests nationwide – and globally.

Addressing bias in AI systems is imperative for risk mitigation. By tackling bias in AI systems throughout the development and management of these systems, businesses can maintain a strong brand reputation and superior value proposition, stay ahead of forthcoming legislation, and be a competitive leader in the fast-paced industry.

Responsible AI systems also can have competitive advantage and gain key market opportunities. Public agencies are increasingly sourcing and using AI systems, and may also be using algorithmic impact assessments. These assessments would prioritize vendors that emphasize fairness, accountability and transparency in their offerings. Responsible AI systems can also improve public trust in the face of high skepticism of societal benefits of tech companies.

BY TACKLING BIAS IN AI SYSTEMS THROUGHOUT THE DEVELOPMENT AND MANAGEMENT OF THESE SYSTEMS, BUSINESSES CAN...

Mitigate risk

Maintain strong brand reputation

Have a superior value proposition

Stay ahead of forthcoming legislation

Be a competitive leader in the fast-paced industry

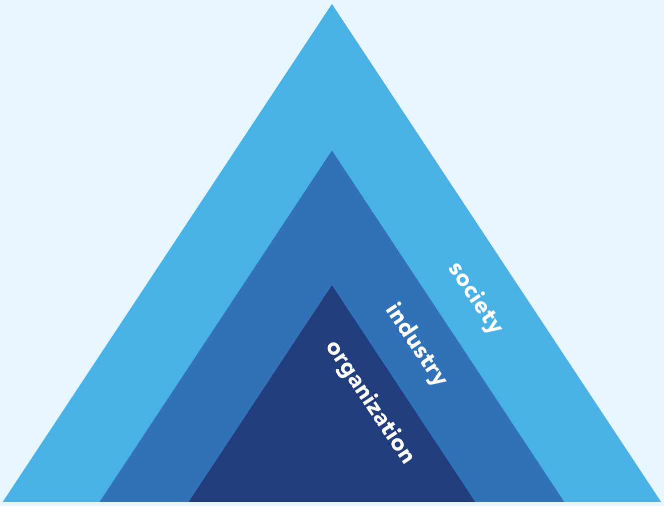
Beware of challenges to mitigate bias

There are various challenges to mitigating biases in AI systems which can be broken down at three levels: organization, industry and society more broadly.

At the **organizational level**, challenges are present in how teams and organizations are set up, as well as engagement with third-party developers and vendors of algorithms.

1 **Lack of diversity in teams:** Tech companies remain primarily male, affluent and white. Challenges related to pipeline are just one part of why there is a lack of diversity in the field, and there are immense issues within technology workplaces—spanning bias in recruiting, hiring, promotions; harassment, violence and sexism; and outdated workplace policies and practices.

2 **Lack of social science and domain knowledge in teams:** Algorithm development teams are typically comprised of data scientists, computer scientists, and/or engineers. Their STEM backgrounds give them a very specific, mathematically aligned approach to analyzing problems—which differs vastly from the way social scientists, philosophers, etc. are taught to address socialproblems.[122] But social science knowledge is critical



to address the various pathways to a biased algorithm and grapple with fairness considerations. While larger companies are increasingly recognizing this, smaller companies and start-ups might not prioritize resources for integrating multidisciplinary knowledge. Even if companies are able to recruit from across disciplines, their retention and promotion criteria are often built with metrics that are primarily compatible with engineering workstreams.[123]

3 “Unknown unknowns”:

The introduction of bias isn’t always evident during a model’s construction—teams might not realize the downstream impacts of their data and choices until much later (especially if they don’t have social science knowledge on their teams). It’s hard to retroactively determine and isolate these “unknown unknowns”. When Amazon attempted to reprogram its hiring AI to ignore words like “women” to stop penalizing female candidates, it found that the algorithm had learned to pick up on implicitly gendered words. They couldn’t fix it, so ended up scrapping the AI altogether.[124] If the test data comes from the same source as the training data, as is often the case, testing will fail to flag skewed / prejudiced results.

4 Focus limited to technical bias and technical solutions:

The term bias is used to say different things. Oftentimes businesses are referring to only technical bias in data without incorporating and addressing other more subtle forms of bias. Solutionst are also technical-focused. These solutions often seek to “de-bias” data in ways that band-aid issues.[125]

5 Limited individual agency and siloed teams:

Looking at individual employees as locus for change remains insufficient[126]—individuals often don’t have the power and/or agency to call out and address ethical issues. Even when they do, different teams are involved in different stages of the development process—and these teams are often siloed.[127] Individuals may not know the specifics of the final product or how their particular work fits in.

6 Lack of accountability:

With various teams and/or third parties working on different components of an algorithm or AI system, internal accountability is not always clear.[128] When working with a third-party vendor, it may be difficult to ensure that their algorithm is unbiased.[129] This is exacerbated by black box models that cannot be explained, further complicating business’ attempts to audit or understand them.

At an **industry-wide level**, AI is a rapidly growing industry where regulatory frameworks are absent or only up-and-coming. Meanwhile, black box algorithms and existing IP laws make it hard to understand algorithm decision making.

1 Market priorities and a fast-paced industry:

Technology companies operate in a fast-paced and disruptive industry where market share and competitive advantage depends on getting new products to the market as soon as possible. Ethical challenges and addressing them takes time. As more and more high-profile incidents of discriminatory AI come to light, risks to business grow. Still, the resulting backlash is often dealt with through reactionary measures like product recalls or other retroactive updates that may not correct the underlying issues.[130]

2 Lack of regulations and actionable guidance:

Businesses are calling for regulatory frameworks and other external governance. While guidelines and parameters from a variety of sources are emerging, they tend to be vague and high-level without actionable tools making it hard to measure progress.[131] International alignment will be critical to making global standards work—but current geopolitical divides hinder the establishment of core values so progress is slow.

3 Persistence of black box algorithms:

The workings of "black box" algorithms are opaque to outside scrutiny. Regulators and consumers are calling for more explainable, "white box" models to be developed to keep the organizations developing AI systems more accountable. But at present, there is a trade-off between simpler models that can be explained and more complex ML models with higher predictive accuracy that have harder to understand feature interactions and inner workings.[132] Improving the accuracy and complexity of explainable models may be possible in future.

4 Machine learning constraints around binary classifications:

In classification ML systems, the algorithm outputs a fixed response to a number of values, often a binary response: yes or no, zero or one ("is this person male or female?"). This presents challenges in trying to capture the gender spectrum. As a result, a simplistic cultural view of gender is built into tools such as image classifications and facial analysis systems, leading to the erasure of nuanced identities that are already marginalized by society. Many researchers acknowledge that gender isn't binary, but may feel limited by constraints of the model. Disability also resists fitting into neat arrangements and fixed classifications with boundaries that continually shift.[133]

5 IP laws constrain understanding algorithm decisions:

Currently, inputs to algorithmic decision-making systems are protected by intellectual property (IP) laws, so it can be difficult to understand how an algorithm was developed and is making decisions. One can only make inferences from outputs as to whether AI systems are expressing bias and why. Governments are starting to address this (e.g., US Congress' Algorithmic Accountability Act would make the code subject to FDA type of review).[134] While not necessarily a challenge for firms that are producing algorithmic systems (as they know what goes into their algorithm decisions), this impacts organizations that purchase or use existing algorithmic systems.

On a broader **societal level**, we are grappling with historical inequities and power dynamics that are reinforced in datasets, teams and decision-making authority, as well as navigating outdated education approaches and "fairness" trade-offs.

1 **Historical inequities, power dynamics and datasets:** Historical inequities and existing power dynamics are pervasive in biased algorithms (see the Bias in AI Map). Also, getting data that is truly representative of different identities for every context is not possible. It is also often fraught with privacy and safety concerns for marginalized communities that are wary of additional surveillance. Meanwhile, certain aspects of identity are simply not considered in ML models (e.g., gender spectrum, (dis)ability).

2 **Lack of diversity in STEM:** The extreme lack of diversity in the field of STEM stymies efforts to enhance diversity in tech companies and reinforces who has power in AI system decision-making.

3 **Outdated education approaches for data and computer scientists:** The lack of ethics, social science and design thinking education in curricula for data and computer scientists[146] limits the ability for individuals working on datasets and algorithms to understand and address potential harms their AI systems.

4 **Legal restrictions:** Current legal structures that mandate race and gender-blind algorithms can actually embed existing inequities by ignoring or hiding it and also make it challenging to check whether discrimination is taking place.

5 **What is "fair?":** Grappling with bias means grappling with notions of "fairness". But there is not one universally accepted definition of "fairness" with various definitions across disciplines. Sometimes these definitions can be at odds with each other. Developers usually attempt to express the concept in mathematical terms, and find their models can only conform to a few "fairness constraints" at a time.[147] Real trade-offs exist between what might be considered "fair" for different groups, and it's hard to determine what definition(s) businesses can / should adopt. Regulations and government engagement have an important role to play here. (See our Fairness Brief for more information and guidance on navigating fairness considerations.)

Execute strategic plays

A. How is this issue being tackled & where does this playbook fit in?

Action is being taken from various stakeholders – spanning companies, academia, government, multilateral institutions, non-governmental organizations (NGOs) and even the Roman Catholic Church[135] to understand and mitigate bias in AI. Academic and industry researchers continue to advance knowledge on fairness, accountability and transparency. Governments, multilateral institutions and NGOs are also contributing to a proliferation of principles, guidelines, and a push for more ethical AI regulations and practices. Many companies are engaging and leading in all aspects, partnering with various stakeholders while working to mitigate biases in AI systems in their own organizations.

Most tech firms are taking initial steps to self-regulate and combat bias. In a survey[136] conducted by DataRobot of 350+ UK- and US-based Chief Information Officers, Chief Technology Officers, and other business leaders involved in AI, 83% of respondents have established AI guidelines and begun to invest in bias prevention initiatives overseen by the C-suite. In large tech companies, this involves establishing senior leadership positions and oversight boards to address AI ethics. These leaders / boards spearhead internal initiatives such as drafting company principles for ethical AI, and may have an external reach such as through supporting relevant academic research. This approach is not without its shortcomings: critics argue that these efforts have not yet gone beyond igniting debates around the social implications of AI[137], often lack operationalization guidance and accountability, and don't explain why certain 'ethical' changes were made.[138] Efforts around responsible AI governance have also been criticized for their lack of transparency

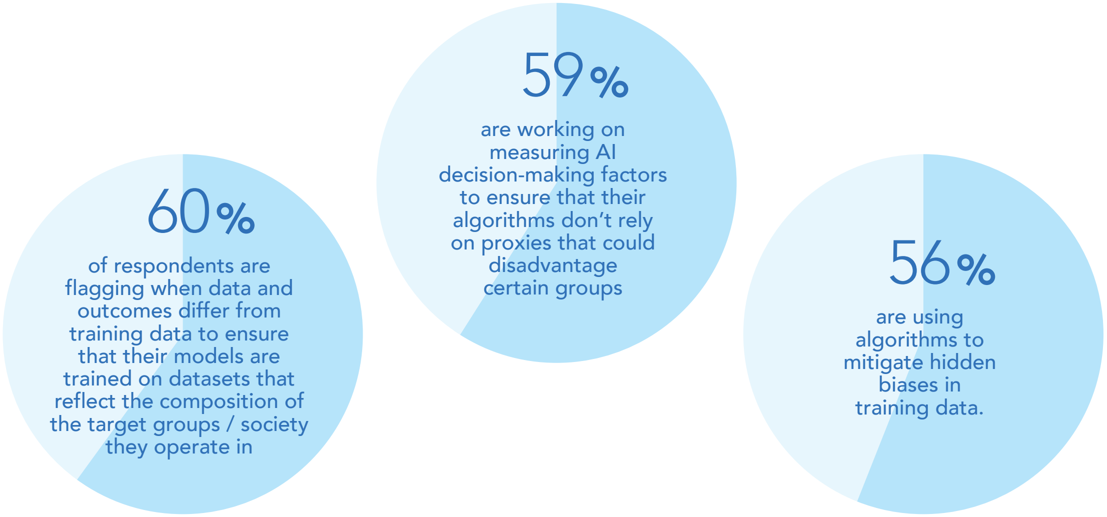


Figure 9. In a DataRobot Survey...

about the criteria used to determine their members, and consequently, whose interests they represent.[139]

At a more granular level, businesses are working to tackle bias creeping in at specific stages in their algorithm development processes. The aforementioned DataRobot survey found that 60% of respondents are flagging when data and outcomes differ from training data to ensure that their models are trained on datasets that reflect the composition of the target groups / society they operate in. Also, 59% are working on measuring AI decision-making factors to ensure that their algorithms don't rely on proxies that could disadvantage certain groups and 56% are using algorithms to mitigate hidden biases in training data (see Figure 9).[140] Businesses are also turning to a range of audits to ensure that their final products are checked for biases and fairness considerations. The industry is seeing a number of third-party algorithmic auditing firms offering these services, with some taking a more technical approach to "de-biasing" AI and others applying a social science lens.[141] Efforts to mitigate bias are only increasing: 93% of surveyed organizations say they will invest more in AI bias prevention initiatives in the next 12 months.

Despite this progress, limitations remain. The majority of efforts focus on addressing "technical" biases through mathematical fixes and fail to address systematic patterns of exclusion as well as power inequities in the field of AI.[142] Some businesses (particularly smaller ones which garner less media attention and have less reputational risk on the line) may not be motivated to address potential sources of bias in their AI at all, eschewing ethical considerations in favor of profit maximization and competitive advantage.

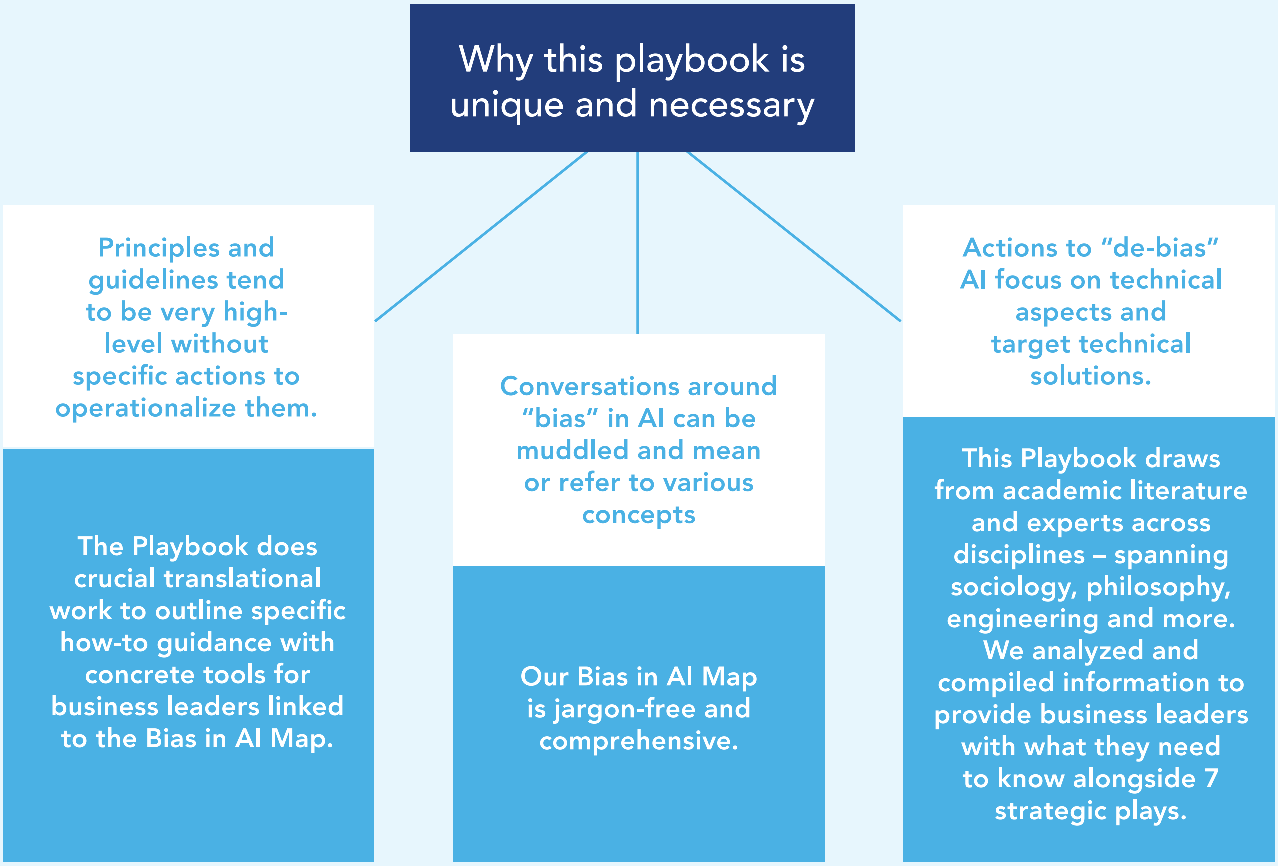
BOX 4. INTEGRATING A HUMAN RIGHTS PERSPECTIVE

While ethics tends to dominate the discourse, human rights is another common lens to understand issues around AI. A human rights approach can draw on universal human rights law and frameworks[148] to understand and highlight issues. For example, biased AI systems can violate human rights such as right to equality and non-discrimination, right to freedom of movement (in the case of surveillance tech) and more. A human rights approach can inform principles for ethical and responsible AI, while also providing guidance for assessing human rights risks in AI and remedying issues.[149]

Recognizing the need for industry-wide change, wider bodies have formed. Partnership on AI, for example, brings together various stakeholders (including think tanks, civil society, academia, international organizations and industry) to promote collaboration across geotechnological lines. Major AI Conferences are making responsible AI an area of focus as well—the 2017 Asilmolar Conference set out 23 principles for beneficial AI, and the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

Governments are establishing working groups to update and introduce new regulations related to responsible AI. The Organisation for Economic Co-operation and Development recently unveiled a non-binding set of principles for AI adopted by 42 countries.[143] The European Commission has 'Ethics Guidelines for Trustworthy AI' and the US government released 10 principles for safer AI in early 2020, with the Department of Defense adopting a series of ethical principles for the use of AI more recently. Likewise, the G20, and Nordic and Baltic states have all published documents related to the development, deployment and uptake of AI. The UN is also delving into the topic, with UNESCO exploring normative frameworks specifically related to gender bias in AI. There is much work to be done to advance regulatory and government action globally. Alongside citizens calling for regulations, are companies – including Google[144] and IBM[145] among others.

This playbook supports companies to act now by filling key knowledge gaps for business leaders on bias in AI and evidence-based solutions to mitigate bias.



Why this playbook is unique and necessary

Principles and guidelines tend to be very high-level without specific actions to operationalize them.

The Playbook does crucial translational work to outline specific how-to guidance with concrete tools for business leaders linked to the Bias in AI Map.

Conversations around "bias" in AI can be muddled and mean or refer to various concepts

Our Bias in AI Map is jargon-free and comprehensive.

Actions to "de-bias" AI focus on technical aspects and target technical solutions.

This Playbook draws from academic literature and experts across disciplines – spanning sociology, philosophy, engineering and more. We analyzed and compiled information to provide business leaders with what they need to know alongside 7 strategic plays.

B. Introducing the plays

Strategic plays for business leaders to mitigate bias in AI span three buckets - teams, AI model, and corporate governance and leadership. We highlight topline elements related to the strategic plays on the following pages.

Strategic plays should be part of your longer term game plan. Some plays also have 'quick wins', which are brief, accompanying resources that can be implemented this quarter and have concrete, immediate benefits.

More information on each of the plays – with how-to guidance, mini cases of leading businesses and tools – and the quick wins can be found separately on the Playbook site. The plays are meant to inspire and inform action plans so customize, use and adapt them as needed.

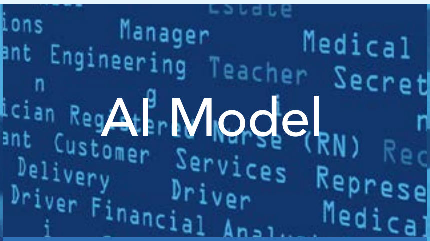
Teams

1

Enable diverse and multi-disciplinary teams working on algorithms and AI systems.

2

Promote a culture of ethics and responsibility related to AI.



AI Model

3

Practice responsible dataset development.

4

Establish policies and practices that enable responsible algorithm development.

Corporate governance & leadership

5

Establish corporate governance for responsible AI and end-to-end internal policies to mitigate bias.

6

Engage corporate social responsibility (CSR) to advance responsible / ethical AI and larger systems change.

7

Use your voice and influence to advance industry change and regulations for responsible AI.

Teams

1. Enable diverse and multi-disciplinary teams working on algorithms and AI systems.

Having diverse teams researching, developing, operationalizing and managing algorithms and AI systems is critical. “Diverse” teams include, but aren’t limited to, women and people of color. Diversity includes individuals with differing personal and group characteristics - such as age, ability, sexual orientation, etc. To have and support diverse teams requires incorporating diversity, equity and inclusion (DEI) principles, policies and practices into the organization, and promoting structural and social changes that enhance diversity in STEM.

Even with diverse teams, we all have blind spots and the onus cannot be simply placed on diverse individuals to identify and mitigate biases. Engaging individuals in the social sciences and humanities – as well as domain experts that understand the particular domain the AI system is meant to operate in – is important. Disciplines including languages, economics, philosophy, psychology and human development include critical philosophical and ethics-based skills that are important for developing and managing AI systems.

ELEMENTS

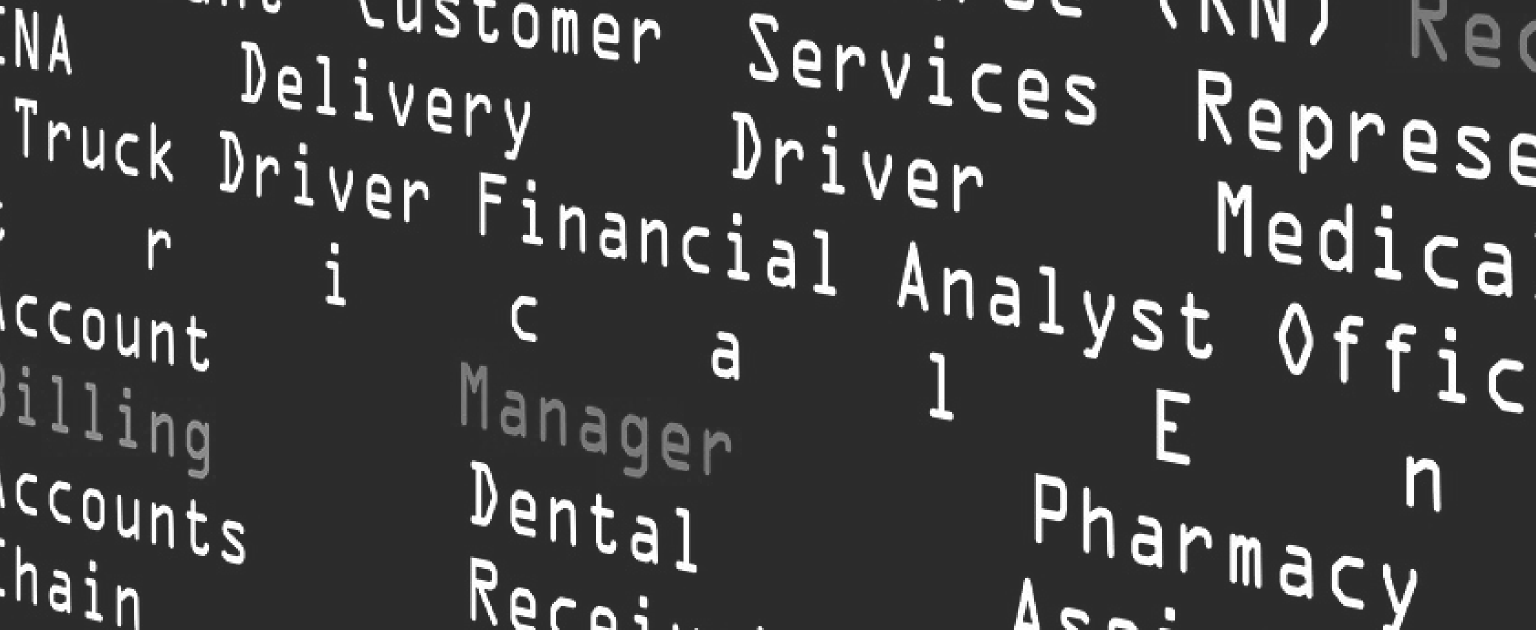
- Ensure diversity is a core leadership priority.
- Update institutional policies, structures and practices to explicitly support diversity and inclusion.
- Ensure teams are multi-disciplinary.
- Measure and share data on diversity metrics, while also reviewing communications.
- Invest in expanding the pipeline for diverse individuals in engineering and data science.
- Invest in research to understand more around advancing DEI in the technology and AI sector.

2. Promote a culture of ethics and responsibility related to AI.

Enable a culture that expects and empowers employees to prioritize equity considerations at every step of the algorithm development process. In line with the understanding that completely de-biasing AI may not be feasible, organizations should uphold a standard of explainability around the workings of their models, as well as transparency around potential shortcomings / pitfalls.

ELEMENTS

- Update individual performance review processes to include a component around responsible and ethical AI practices.
- Update Objectives & Key Results (OKRs) / Key Performance Indicators (KPIs) to integrate goals and metrics on mitigation of bias in AI.
- Embed training on ethics, bias and fairness for employees developing, managing and/or using AI systems.
- Embed training on language related to diversity, equity and inclusion for employees and/or contractors labeling data.
- Make explainability and transparency around shortcomings and pitfalls of AI systems the norm.



AI Model

3. Practice responsible dataset development.

Ensure that dataset development is conducted responsibly, with standard checks and balances in place for creating new datasets as well as adapting existing ones. The creation and implementation of such practices requires businesses to be intentional about gathering inclusive data and asking important questions around who is benefiting from the data collected.

ELEMENTS

- Use broader training sets to remove sample / selection bias and ensure that various identities are sufficiently and accurately represented.
- Adopt a "data with depth" approach that integrates social science methods into more technical data science methods for generating and collecting data.
- Place checks on labeling practices.
- Document the provenance, creation, and use of ML datasets; ensure that the purpose of the dataset aligns with its intended use.
- Assess existing datasets to check for over-/ under-representation of certain identities, underlying inequities that reflect reality but are ultimately problematic, and address privacy concerns.
- Maintain datasets as living resources.

4. Establish policies and practices that enable responsible algorithm development.

Build practices that check for and actively mitigate bias into every stage of the algorithm development process. This involves equipping teams with ethical frameworks that allow them to prioritize equity while defining their algorithms' objectives, ensuring datasets used are responsibly developed and labeled, and ensuring variables do not disadvantage certain communities.

ELEMENTS

- Establish an ethical framework within which the AI system's purpose and objectives are defined.
- Ensure that datasets and proxies chosen do not advertently / inadvertently disadvantage certain identities.
- Document the provenance and development of the AI system – including data sources and variables.
- Integrate human-in-the-loop processes.
- Engage communities impacted by the AI systems in their development, where possible.
- Conduct internal and external audits on the AI system.

Corporate governance & leadership

5. Establish corporate governance for responsible AI and end-to-end internal policies to mitigate bias.

It is important to establish corporate governance for responsible AI and end-to-end internal policies and guidance to mitigate bias. AI ethics governance structures is a first step. Good practices for responsible AI governance include, for example: cultivating a sense of shared responsibility, assessing and updating incentive structures and power dynamics that can dissuade individuals from speaking up, and examining leadership priorities and limitations.

ELEMENTS

- Establish an AI ethics / responsible AI lead, an AI ethics board and AI ethics code / principles.
- Establish and formalize processes including concrete guidance and tools to help plan for, identify and mitigate biases in AI systems.
- Assess how leadership priorities can impact responsible AI practices. Be honest, as well as transparent, about limitations.

6. Engage corporate social responsibility (CSR) to advance responsible / ethical AI and systems change.

AI systems and their biases mirror and replicate existing structures of power and inequality in society. While it is not necessarily the responsibility of the business to correct for or adjust these inequalities, CSR and/or corporate foundations can promote equity and inclusion in material, strategic areas. CSR teams can serve as an incubator for different interventions internally while also supporting or developing longer-term interventions externally. CSR teams are well placed for this as they operate under a different incentive structure than technical teams, which tend to prioritize efficiency. Beyond advancing internal efforts, CSR teams can also be deployed to address biases in data; address power dynamics and lack of diversity in AI; and catalyze research and education (for data scientists, engineers and business students) on responsible AI.

ELEMENTS

- Leverage CSR teams to advance internal bias mitigation efforts.
- Align CSR efforts to mitigate bias in AI and support long term systems change with the company's own goals and material, strategic interests.

7. Use your voice and influence to advance industry change and regulations for responsible AI.

The field of AI is fast-moving and it's critical that business leaders stay up-to-date in understandings around bias in AI systems. Leaders must acknowledge that bias in AI systems is not a purely technical issue, but also linked to societal inequities. These inequities are not just mirrored and reinforced in biased datasets, but also in the people and systems that are creating, generating and collecting data as well as designing, developing and operating AI systems. A responsible business leader understands this larger, holistic view and the trade-offs related to "fairness" that can be at play. Business leaders can use their voice and influence to support industry change and needed regulations.

ELEMENTS

- Join / initiate meaningful partnerships with other companies, governments, academics and/or non-profit organizations. This may involve starting or joining working groups or industry associations to inform or advocate for responsible public policies to govern AI and approaches for industry.
- Contribute to ongoing debates around bias in AI and insist on / support meaningful dialogue among wider array of stakeholders in algorithmic accountability – such as vulnerable communities impacted by technology, nonprofits on the frontlines fighting discrimination and injustice, legislators and regulators.
- Fund research to advance knowledge in the space of responsible AI (especially diverse research teams) and prioritize working with other organizations or initiatives that have diverse teams and/or responsible data / AI systems practices.

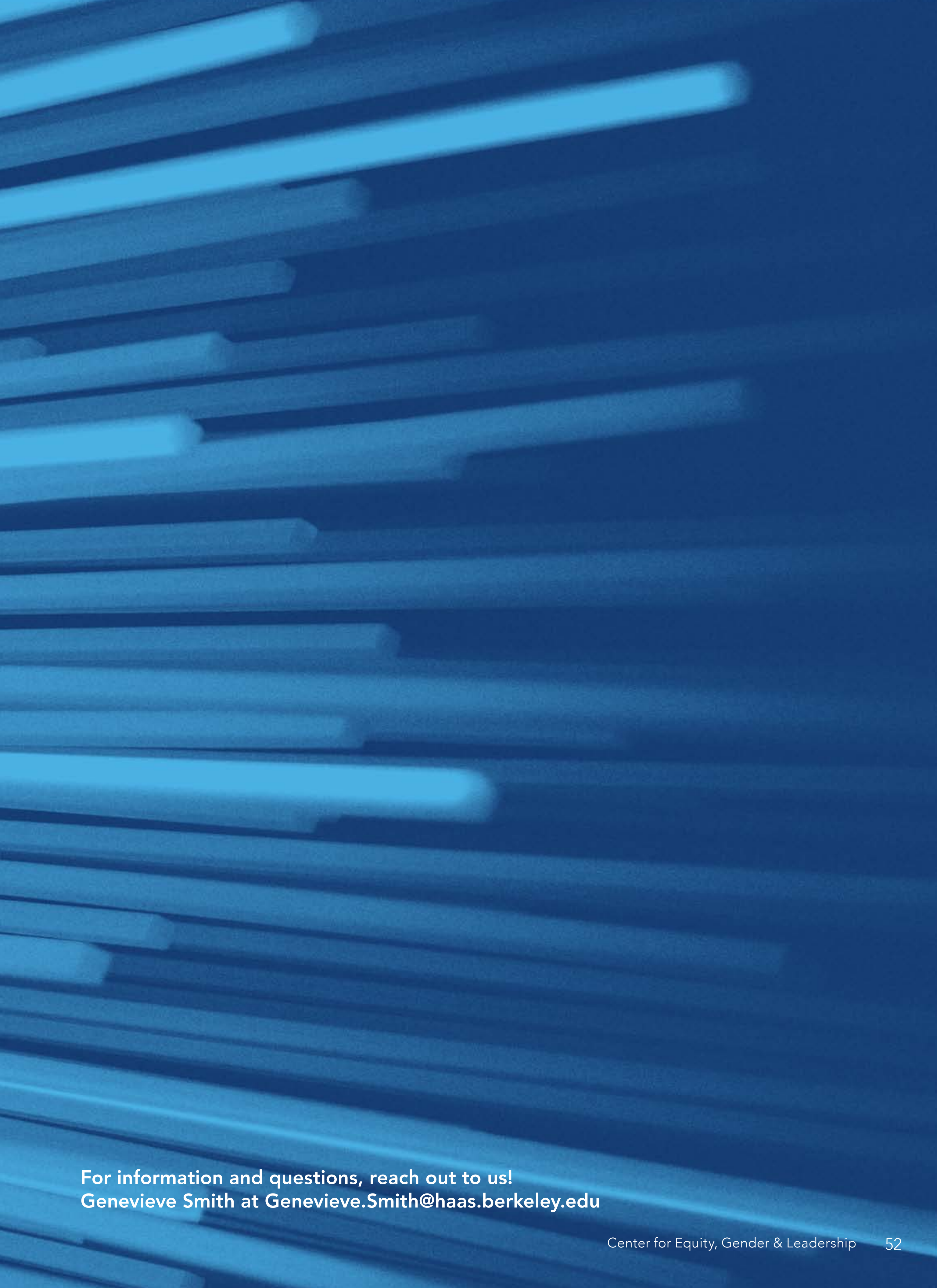
Call to Action

Remember Anita? She later worked with a new team of developers at her firm alongside diversity and inclusion experts to map out how bias could creep into the dataset, what proxies used in the algorithm might inadvertently lead to bias and an evaluation process to understand unintended bias in the algorithm. The team also set out to develop an AI system with a different purpose in mind from the get-go: to equitably identify top candidates for the firm that could add value to the firm and lead to a diverse, thriving organization. It is still a work in progress, but has had early success.

By understanding how bias in AI can manifest and implementing the seven plays, business leaders can unlock the incredible potential of AI in ways that advance the business and society. They can mitigate risk and be the captains of businesses at the forefront.

Our goal with this Playbook is to guide you and your organization towards concrete, meaningful action. We now turn it over to you, to take the playbook and mitigate bias in AI to unlock its transformative value responsibly and equitably. As you go along your journey, reach out to us and share your challenges and lessons learned so we can continue to track and update learnings and approaches.

Yours in learning and equity,
The EGAL Team



For information and questions, reach out to us!
Genevieve Smith at Genevieve.Smith@haas.berkeley.edu

Glossary

Algorithm: Mathematical instructions that give instructions to computers. Algorithms are designed by humans and used in AI systems to make decisions using data.

Artificial intelligence (AI): Machines that respond to stimulation consistent with traditional responses from humans, given the human capacity for contemplation, judgment and intention. More simply, AI implies the use of a computer to model and/or replicate intelligent behavior.

Audit: A tool for interrogating complex processes to determine whether processes are compliant with company policy, industry standards or regulations. While algorithmic audits are new, auditing is a common practice and its own industry. Audits are generally concerned not only with output of a specific system – but also the process of checks, control and quality of the system itself.[150]

Bias: A tendency, inclination, or prejudice toward or against something or someone. Biases are often based on stereotypes, rather than actual knowledge of an individual or circumstance. These cognitive shortcuts can result in prejudgments that can lead to discriminatory practices.In AI and ML, the term "bias" is often used and defined in a technical sense. However, the broader concept of bias considered in this playbook illustrates how many forms of discrimination can emerge from AI systems -- even when AI systems are working as intended.

Biased AI: This playbook refers to AI systems that result in either (1) incorrect outputs / predictions for certain populations and/or (2) discriminatory output / predictions for certain populations as "biased AI". Biased AI systems can unfairly allocate opportunities, resources or information; infringe on civil liberties; pose a detriment to the safety of individuals; fail to provide the same quality of service to some people as others; and negatively impact a person's wellbeing such as by being derogatory or offensive.

Black box model: AI system for which users are able to observe inputs and outputs, but are unable to follow the exact decision making process. Machine learning algorithms internalize massive amounts of data, and instead of storing what they have learnt in a neat block of digital memory, they diffuse the information in a way that is exceedingly difficult to decipher. These factors make such complex systems opaque – often even to their creators.

Data: Data encompasses a lot – it can be numbers, words, images, clicks, etc. Vast amounts of data points are generated by virtue of individuals' day-to-day activities (e.g., consumer behavior, health conditions) and data points are collected through various platforms, technological or otherwise.

Dataset development: The acquisition, cleaning, and labeling of large amounts of data points to prepare datasets for algorithms to (1) learn from, (2) test their learnings against, and (3) perform their operations on.

Deep learning: A subset of ML in which models make their own predictions entirely independent of humans (after the models are created). DL structures algorithms in layers to create artificial neural networks (inspired by the biological neural network of the human brain) and makes it possible for the computer program to learn on its own.

Diversity: The wide variety of shared and different personal and group characteristics among human beings (including but not limited to Race, Ethnicity, Gender, Age, Ability, Religion, Sexual Orientation, Socio-economic Status, etc.).

Equity: The process of being treated fairly or impartially. Specific definitions vary across contexts, and this concept is tied closely to notions of fairness and ethics.

Ethics: Commonly defined as a set of moral issues or aspects. The lack of a more specific, universally accepted definition makes this a difficult concept to operationalize, particularly in context of mathematical AI models. It is closely tied to the notions of equity and fairness.

Fairness: Commonly defined as the quality or state of being fair, especially fair or impartial treatment. But what exactly fairness means in the context of ML is not clear. Fairness is a complex concept and deeply contextual. Also, various disciplines conceptualize "fairness" differently. Read more on how different disciplines define fairness and what fairness means in the context of ML **here**.

Machine learning (ML): A type of AI that is made up of a series of algorithms and takes / learns from massive amounts of data to find patterns and make predictions. It performs a function with the data given to it and gets progressively better over time.

Proxy: The variables that machine learning algorithms use to predict outcomes. More specifically, proxies are used instead of the variable of interest when that variable of interest cannot be measured directly. For example, per capita GDP can be used as a proxy for the standard of living.[151]

"Reinforced" machine learning: A reinforcement algorithm learns by trial and error to achieve a clear objective (e.g., AlphaGo).[152]

"Supervised" machine learning: The most prevalent form of ML, in which data is labeled to tell the algorithm exactly what patterns to look for.[153]

Training dataset: The input data used by a machine learning algorithm to find patterns. For machine learning that examines images or videos, training data will include those images or videos themselves. Training data is generally made up of variables and a "target variable" or "training label" that a machine learning model will attempt to predict.[154]

"Unsupervised" machine learning: Machine learning systems in which the data hasno labels and the algorithm looks for whatever patterns it can find.[155]

White box model: AI model that satisfies two key criteria: its features (variables) are understandable, and its decision-making process is clear and explainable.[156] This often refers to AI systems with less predictive capacity than machine learning, such as those that use linear regression or decision trees. These models are significantly easier to explain and interpret, but may not always be capable of modelling the inherent complexity of their datasets.[157]

Endnotes

1 Sizing the prize. PwC. Retrieved on March 24, 2020 from https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html.

2 (2020). From roadblock to scale: The global sprint towards AI. Morning Consult & IBM. Retrieved from http://filecache.mediaroom.com/mr5mr_ibmnews/183710/Roadblock-to-Scale-exec-summary.pdf.

3 (2018). United States Securities and Exchange Commission Form 10-Q: Microsoft Corporation. Retrieved on March 23, 2020 from https://www.sec.gov/Archives/edgar/data/789019/000156459019001392/msft-10q_20181231.htm.

4 Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

5 West, S. M., Whittaker, M. & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. AI Now Institute. Retrieved from https://ainowinstitute.org/discriminatingsystems.pdf.

6 Collett, C. & Dillon, S. (2019). AI and gender: Four proposals for future research. University of Cambridge & Leverhulme Centre for the Future of Intelligence. Retrieved from http://lcfi.ac.uk/media/uploads/files/AI_and_Gender___4_Proposals_for_Future_Research.pdf.

7 West, S. M., Whittaker, M. & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. AI Now Institute. Retrieved from https://ainowinstitute.org/discriminatingsystems.pdf.

8 Ashcraft, C., McLain, B. & Eger, E. (2016). Women in tech: The facts. National Center for Women in Information Technology. Retrieved from https://www.ncwit.org/sites/default/files/resources/womenintech_facts_fullreport_05132016.pdf.

9 Global diversity and inclusion. Microsoft. Retrieved on March 19, 2020 from https://www.microsoft.com/en-us/diversity/inside-microsoft/default.aspx.

10 Crawford, K. (2013). Hidden biases in big data. HBR.

11 Benjamin, R. (2019). Race after technology. Polity Press.

12 Edmundson, A. (2020, January 24). Getting oriented in the datafied world – thinking in time [Lecture notes].

13 The State of AI Bias in 2019. (n.d.). Retrieved from https://www.datarobot.com/lp/the-state-of-ai-bias-in-2019/

14 Crawford, K. & Paglen, T. (2019). Excavating AI: The politics of images in machine learning training sets. Retrieved from https://www.excavating.ai/.

15 Obermeyer, Z., Powers, B. Vogeli, C., Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366 (6464), 447-453.

16 The Roman Catholic Church joined with IBM and Microsoft to work on ethics of artificial intelligence in the "Rome Call for Ethics". The call, which outlines three principles, was supported by Pope Francis who made detailed remarks about the impact of AI on humanity. Read more: https://romecall.org/.

17 (2019). 22nd annual global CEO survey. PwC. Retrieved from https://www.pwc.com/mu/pwc-22nd-annual-global-ceo-survey-mu.pdf.

18 Sizing the prize. PwC. Retrieved on March 24, 2020 from https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html.

19 Cohen, R. & Bray, D. (2020). COVID-19 might accelerate or change previous AI adoption strategies. Atlantic Council. Retrieved from https://atlanticcouncil.org/blogs/geotech-cues/covid-19-might-accelerate-or-change-previous-ai-adoption-strategies/.

20 Shubhendu, S., & Vijay, F. (2013). Applicability of artificial intelligence in different fields of life. International Journal of Scientific Engineering and Research, 1 (1).

21 West, D. (2018). What is artificial intelligence? Brookings. Retrieved from https://www.brookings.edu/research/what-is-artificial-intelligence/.

22 Wu, J. (2019). AI, machine learning, deep learning explained simply. Towards Data Science. Retrieved from https://towardsdatascience.com/ai-machine-learning-deep-learning-explained-simply-7b553da5b960.

23 Wu, J. (2019). AI, machine learning, deep learning explained simply. Towards Data Science. Retrieved from https://towardsdatascience.com/ai-machine-learning-deep-learning-explained-simply-7b553da5b960.

24 Hao, K. (2018). What is machine learning. MIT Technology Review. Retrieved from https://www.technologyreview.com/s/612437/what-is-machine-learning-we-drew-you-another-flowchart/.

25 Artificial intelligence: What it is and why it matters. Sas. Retrieved on March 22, 2020 from https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html.

26 Edmundson, A. (2020, January 27). How the world was datafied – historical overview and timeline [Lecture notes].

27 (2020). From roadblock to scale: The global sprint towards AI. Morning Consult & IBM. Retrieved from http://filecache.mediaroom.com/mr5mr_ibmnews/183710/Roadblock-to-Scale-exec-summary.pdf.

28 (2020). From roadblock to scale: The global sprint towards AI. Morning Consult & IBM. Retrieved from http://filecache.mediaroom.com/mr5mr_ibmnews/183710/Roadblock-to-Scale-exec-summary.pdf.

29 (2018). Machine learning survey. UNIVA. Retrieved from https://www.univa.com/resources/univa-machine-learning-survey.php.

30 (2020). From roadblock to scale: The global sprint towards AI. Morning Consult & IBM. Retrieved from http://filecache.mediaroom.com/mr5mr_ibmnews/183710/Roadblock-to-Scale-exec-summary.pdf.

31 Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association : JAMIA, 19(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

32 Manyika, J., Silberg, J. & Presten, B. (2019). What do we know about biases in AI? HBR.

33 Digital assistants tend to be feminized and both reflect as well as amplify troubling gender biases. For more information on gendered voice assistants and this issue, see: I'd Blush if I Could (UNESCO).

34 Bias. (n.d.). Retrieved from https://www.psychologytoday.com/us/basics/bias.

35 Kahneman, D. (2019). Daniel Kahneman Explains The Machinery of Thought. Retrieved from https://fs.blog/2014/07/daniel-kahneman-the-two-systems/.

36 Eberhardt, J. (2020). Biased. Penguin Random House.

37 Umoja Noble, S. (2018). Algorithms of oppression: How search engines reinforce racism. New York

University Press.

38 Collett, C. & Dillon, S. (2019). AI and gender: Four proposals for future research. University of Cambridge & Leverhulme Centre for the Future of Intelligence. Retrieved from http://lcfi.ac.uk/media/uploads/files/AI_and_Gender___4_Proposals_for_Future_Research.pdf.

39 West, S. M., Whittaker, M. & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. AI Now Institute. Retrieved from https://ainowinstitute.org/discriminatingsystems.pdf.

40 Ashcraft, C., McLain, B. & Eger, E. (2016). Women in tech: The facts. National Center for Women in Information Technology. Retrieved from https://www.ncwit.org/sites/default/files/resources/womenintech_facts_fullreport_05132016.pdf.

41 (2019). Global AI talent report 2019. Element AI. Retrieved from https://jfgagne.ai/talent-2019/.

42 (2018). Artificial Intelligence index 2018. AI Index 2018. Retrieved from http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf

43 West, S. M., Whittaker, M. & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. AI Now Institute. Retrieved from https://ainowinstitute.org/discriminatingsystems.pdf.

44 Global diversity and inclusion. Microsoft. Retrieved on March 19, 2020 from https://www.microsoft.com/en-us/diversity/inside-microsoft/default.aspx.

45 Crawford, K. (2013). Hidden biases in big data. HBR.

46 Benjamin, R. (2019). Race after technology. Polity Press.

47 Edmundson, A. (2020, January 24). Getting oriented in the datafied world – thinking in time [Lecture notes].

48 Collett, C. & Dillon, S. (2019). AI and gender: Four proposals for future research. University of Cambridge & Leverhulme Centre for the Future of Intelligence. Retrieved from http://lcfi.ac.uk/media/uploads/files/AI_and_Gender___4_Proposals_for_Future_Research.pdf.

49 Neff, G. (2020, February 19). Personal interview.

50 Williams, A., Brooks, C. & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions and policy implications. Journal of Information Policy, 8, 78-115.

51 (2017). A call to action for gender equality and women's economic empowerment. UN High Level Panel on Women's Economic Empowerment.

52 Mcfarling, U. (2020). Fearing deportation, many immigrants at higher risk of Covid-19 are afraid to seek testing or care. Stat. Retrieved from https://www.statnews.com/2020/04/15/fearing-deportation-many-immigrants-at-higher-risk-of-covid-19-are-afraid-to-seek-testing-or-care/.

53 Retrieved from Corbyn, Z. (2020). Interview: Catherine D'Ignazio. The Guardian.

54 (2018). Clinical trials have far too little racial and ethnic diversity. Scientific American. Retrieved from https://www.scientificamerican.com/article/clinical-trials-have-far-too-little-racial-and-ethnic-diversity/.

55 2001). Drug safety: Most drugs withdrawn in recent years had greater health risks for women. United States Senate. Retrieved from https://www.gao.gov/assets/100/90642.pdf.

56 Mikhail GW. Coronary heart disease in women. BMJ. 2005;331(7515):467-468. doi:10.1136/bmj.331.7515.467

57 (2018). Clinical trials have far too little racial and ethnic diversity. Scientific American. Retrieved from https://www.scientificamerican.com/article/clinical-trials-have-far-too-little-racial-and-ethnic-diversity/.

58 (2018). Clinical trials have far too little racial and ethnic diversity. Scientific American. Retrieved from https://www.scientificamerican.com/article/clinical-trials-have-far-too-little-racial-and-ethnic-diversity/.

59 (2019). Genetics for all. Nature Genetics 51, 579. https://doi.org/10.1038/s41588-019-0394-y.

60 Edmundson, A. (2020, January 27). How the world was datafied – historical overview and timeline [Lecture notes].

61 Perez, C. C. (2019). Invisible women. Abrams Press.

62 (2020). Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed Coronavirus disease 2019. CDC. Retrieved from https://www.cdc.gov/mmwr/volumes/69/wr/mm6915e3.htm?s_cid=mm6915e3_w.

63 Zeng, F., Dai, C., Cai, P., Wang, J., Xu, L., Li, J., Hu, G. & Wang, L. (2020). A comparison study of SARS-CoV-2 IgG antibody between male and female COVID-19 patients. medRxiv. https://doi.org/10.1101/2020.03.26.20040709.

64 Collett, C. & Dillon, S. (2019). AI and gender: Four proposals for future research. University of Cambridge & Leverhulme Centre for the Future of Intelligence. Retrieved from http://lcfi.ac.uk/media/uploads/files/AI_and_Gender___4_Proposals_for_Future_Research.pdf.

65 Silberg, J. & Manyika, J. (2019). Notes from the AI frontier: Tackling bias in AI (and humans). McKinsey Global Institute. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/MGI-Tackling-bias-in-AI-June-2019.ashx.

66 Sampson, R., & Lauritsen, H. (1997). Racial and ethnic disparities in crime and criminal justice in the United States. Crime and Justice, 21: 311-374. https://doi.org/10.1086/449253

67 Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16). doi: 10.1073/pnas.1720347115

68 Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems, 4349-4357.

69 Latanya Sweeney, "Discrimination in Online Ad Delivery," Queue, March 2013, Volume 11, Issue 3

73 Lief, L. (2020). How philanthropy can help lead on data justice. Stanford Social Innovation Review. Retrieved from https://ssir.org/articles/entry/how_philanthropy_can_help_lead_on_data_justice.

70 Buolamwini, J. & Gebru, T. (2018). Gender shares: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81: 1-15.

71 Puri, R., (2018). Mitigating bias in AI models. IBM. Retrieved from https://www.ibm.com/blogs/research/2018/02/mitigating-bias-ai-models/.

72 Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. doi: 10.1038/nature21056

74 https://www.excavating.ai/

75 Machine Learning: Analyzing Gender. (n.d.). Retrieved from http://genderedinnovations.stanford.edu/case-studies/machinelearning.html#tabs-2

76 Whittaker, M., Alper, M., Bennett, C., Hendren, S., Kaziunas, L., Mills, M., Morris, M. R., Rankin, J., Rogers, E., Salas, M., West, S. M. (2019). Disability, power and AI. AI Now.

77 West, S. (2020, February 12) Personal Interview.

78 Benjamin, R. (2019). Race after technology. Polity Press.

79 Zhou, J. (2020, January 23). Personal interview.

80 Russell, S. (2020, February 5). Personal interview.

81 Kleinberg, J., Ludwig, J., Mullainathan, S. & Rambachan, A. (2018). Advances in big data research in economics: Algorithmic fairness. AEA Papers and Proceedings 2018, 108, 22-27. https://doi.org/10.1257/pandp.20181018.

82 Spielkamp, M. (2017). Inspecting algorithms for bias. MIT Technology Review. Retrieved from https://www.technologyreview.com/2017/06/12/105804/inspecting-algorithms-for-bias/.

83 Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

84 Proxies are defined as a variable used instead of the variable of interest when that variable of interest cannot be measured directly. For example, per capita GDP can be used as a proxy for the standard of living. Retrieved from Oxford Reference at https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100351624.

85 Edwards, H. The quest to make AI less prejudiced. Quartz. Retrieved from https://qz.com/1814415/how-to-fix-bias-in-ai/.

86 Terrell, J., Kofnik, A., Middleton, J., Rainear, C., Murphy-Hill, E., Parnin, C. & Stallings, J. (2016). Gender differences and bias in open source: Pull request acceptance of women versus men. PeerJ Preprints 4:e1733v2 https://doi.org/10.7287/peerj.preprints.1733v2

87 Perez, C. C. (2019). Invisible women. Abrams Press.

88 Obermeyer, Z., Powers, B. Vogeli, C., Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366 (6464), 447-453.

89 Williams, B., Brooks, C. & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions and policy implications. Journal of Information Policy, 8, 78-115.

90 Kleinberg, J., Ludwig, J., Mullainathan, S. & Rambachan, A. (2018). Advances in big data research in economics: Algorithmic fairness. AEA Papers and Proceedings 2018, 108, 22-27. https://doi.org/10.1257/pandp.20181018.

91 Mulligan, D., Elazari, A., Burrell, J, & Kluttz, D., (2018). AFOG workshop panel 2: Automated decision-making is imperfect, but it's arguably an improvement over biased human decision-making. UC Berkeley AFOG. Retrieved from https://afog.berkeley.edu/files/2019/07/AFOG_workshop2018_report_all_web.pdf.

92 Kleinberg, J., Ludwig, J., Mullainathan, S. & Rambachan, A. (2018). Advances in big data research in economics: Algorithmic fairness. AEA Papers and Proceedings 2018, 108, 22-27. https://doi.org/10.1257/pandp.20181018.

93 Mulligan, D., Elazari, A., Burrell, J, & Kluttz, D., (2018). AFOG workshop panel 2: Automated decision-making is imperfect, but it's arguably an improvement over biased human decision-making. UC Berkeley AFOG. Retrieved from https://afog.berkeley.edu/files/2019/07/AFOG_workshop2018_report_all_web.pdf.

94 Colson, E. (2020, April 2). What AI-Driven Decision Making Looks Like. Retrieved from https://hbr.org/2019/07/what-ai-driven-decision-making-looks-like

95 Shema, Alain. (2019). Effective credit scoring using limited mobile phone data. 1-11. 10.1145/3287098.3287116.

96 Gertler, P. (2020, February 18). Personal interview.

97 The Clooney Foundation for Justice. (2019). What is TrialWatch? Retrieved from https://cfj.org/project/trialwatch/

98 SEER Cancer Statistics Review, 1975-2004 (2019). Ethnicity and Skin Cancer. Retrieved from https://dermlite.com/pages/ethnicity

99 IBM Research Blog. (2016). Identifying Skin Cancer with Computer Vision. Retrieved from https://www.ibm.com/blogs/research/2016/11/identifying-skin-cancer-computer-vision/.

100 Cowgill, B. (2019). Bias and Productivity in Humans and Machines. SSRN Electronic Journal. doi: 10.2139/ssrn.3433737

101 Benjamin, R. (2019). Race after technology. Polity Press.

102 This was informed by a brief developed by Matt McGee in November 2019 (UC Berkeley, Haas School of Business, MBA Candidate)

103 Friedman, B., & Nissenbaum, H. (2017). Bias in Computer Systems. Computer Ethics, 215–232. doi: 10.4324/9781315259697-23

104 Hill, K. (2020). Before Clearview became a police tool, it was a secret plaything of the rich. NY Times. Retrieved from https://www.nytimes.com/2020/03/05/technology/clearview-investors.html.

105 Hill, K. (2020, January 18). The Secretive Company That Might End Privacy as We Know It. Retrieved from https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

106 Barbaschow, A. (2020, May 28). ACLU sues Clearview AI claiming the company's tech crosses ethical bounds. Retrieved from https://www.zdnet.com/article/aclu-sues-clearview-ai-claiming-the-companys-tech-crosses-ethical-bounds/

107 Biddle, S. (2020). ICE's New York offices uses rigged algorithm to keep virtually all arrestees in detention. The ACLU says its unconstitutional. The Intercept. Retrieved from https://theintercept.com/2020/03/02/ice-algorithm-bias-detention-aclu-lawsuit/.

109 Madaio, M., Stark, L., Vaughan, J. W. & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. CHI 2020 paper. http://www.jennwv.com/papers/checklists.pdf.

110 This is a key feature of the "New Jim Code", as outlined by Ruha Benjamin in her book Race After Technology, in which coded inequity makes discrimination easier, faster and even harder to challenge. Source: Benjamin, R. (2019). Race after technology. Polity Press.

111 https://www.csee.umbc.edu/~cmat/Pubs/KayMatuszekMunsonCHI2015GenderImageSearch.pdf

112 Perez, C. C. (2019). Invisible women. Abrams Press.

113 (2018). United States Securities and Exchange Commission Form 10-Q: Microsoft Corporation. Retrieved on March 23, 2020 from https://www.sec.gov/Archives/edgar/data/789019/000156459019001392/msft-10q_20181231.htm.

114 The State of AI Bias in 2019. (n.d.). Retrieved from https://www.datarobot.com/lp/the-state-of-ai-bias-in-2019/

115 2019 Edelman Trust Barometer. (n.d.). Retrieved from https://www.edelman.com/research/2019-edelman-

trust-barometer

116 A survey by MetLife found that 89% of employees would trade some of their salary to work at a company whose values match their own. Source: (2017). Employees to employers: We want you to share our values and make the world a better place. MetLife. Retrieved from https://www.metlife.com/about-us/newsroom/2017/november/employees-to-employers–we-want-you-to-share-our-values-and-make/.

117 Shane, S., & Wakabayashi, D. (2018, April 4). 'The Business of War': Google Employees Protest Work for the Pentagon. Retrieved from https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html

118 Daws, R. (2020, March 25). Microsoft warns its AI offerings 'may result in reputational harm'. Retrieved from https://artificialintelligence-news.com/2019/02/06/microsoft-ai-result-reputational-harm/

119 Hao, K. (2020, June 15). The two-year fight to stop Amazon from selling face recognition to the police. Retrieved from https://www.technologyreview.com/2020/06/12/1003482/amazon-stopped-selling-police-face-recognition-fight/

120 Marantz, A., Max, D. T., Schwartz, A., & Donohue, J. (n.d.). When an App Is Called Racist. Retrieved from https://www.newyorker.com/business/currency/what-to-do-when-your-app-is-racist

121 Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. The Nation.

108 The State of AI Bias in 2019. (n.d.). Retrieved from https://www.datarobot.com/lp/the-state-of-ai-bias-in-2019/

122 Hao, K. (2020, April 2). This is how AI bias really happens-and why it's so hard to fix. Retrieved from https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/?utm_source=newsletters&utm_medium=email&utm_campaign=the_algorithm.unpaid.engagement

123 Ammanath, B. (2020, April 4). Personal interview.

124 Hao, K. (2020, April 2). This is how AI bias really happens-and why it's so hard to fix. Retrieved from https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/?utm_source=newsletters&utm_medium=email&utm_campaign=the_algorithm.unpaid.engagement

125 Zhou, J. (2020, January 23).. Personal interview.

126 West, S. (2020, February 12) Personal interview.

127 Machine Learning: Analyzing Gender. (n.d.). Retrieved from http://genderedinnovations.stanford.edu/case-studies/machinelearning.html#tabs-2

128 Benjamin, R. (2019). Race after technology. Polity Press.

129 Finlayson, J. (2020, February 5). Personal interview.

130 Leetaru, K. (2019, January 21). Why Is AI And Machine Learning So Biased? The Answer Is Simple Economics. Retrieved from https://www.forbes.com/sites/kalevleetaru/2019/01/20/why-is-ai-and-machine-learning-so-biased-the-answer-is-simple-economics/#51eb3979588c

131 Pichai, S. (2020, January 20). Why Google thinks we need to regulate AI. Retrieved from https://www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04

132 Hulstaert, L. (2019, March 14). Machine learning interpretability techniques. Retrieved from https://towardsdatascience.com/machine-learning-interpretability-techniques-662c723454f3

133 Whittaker, M., Alper, M., Bennett, C., Hendren, S., Kaziunas, L., Mills, M., Morris, M. R., Rankin, J., Rogers, E., Salas, M., West, S. M. (2019). Disability, power and AI. AI Now.

134 Nkonde, M. (2019). Is AI bias a corporate social responsibility issue? HBR. Retrieved from https://hbr.org/2019/11/is-ai-bias-a-corporate-social-responsibility-issue.

135 The Roman Catholic Church joined with IBM and Microsoft to work on ethics of artificial intelligence in the "Rome Call for Ethics". The call, which outlines three principles, was supported by Pope Francis who made detailed remarks about the impact of AI on humanity. Read more: https://romecall.org/.

136 The State of AI Bias in 2019. (n.d.). Retrieved from https://www.datarobot.com/lp/the-state-of-ai-bias-in-2019/

137 https://www.ft.com/content/a3328ce4-60ef-11e9-b285-3acd5d43599e

138 For instance, when Microsoft claimed it cut off "significant sales" as per its AI ethics committee's recommendations, it never explained what was determined as unethical, and by what process, making it impossible to discern the board's oversight capabilities. https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech

139 Levin, S. (2019, April 5). Google scraps AI ethics council after backlash: 'Back to the drawing board'. Retrieved from https://www.theguardian.com/technology/2019/apr/04/google-ai-ethics-council-backlash

140 A range of algorithms can proactively parse through training datasets to look for over-/underrepresentation of certain communities, skewed word associations. But other forms of biases that can creep into data during collection or labeling processes (see the Bias in AI Map for a more comprehensive understanding of how biases manifest in datasets). Major tech companies make such algorithms open source and available for use. https://analyticsindiamag.com/top-5-tools-data-scientists-can-use-to-mitigate-biases-in-algorithms/

141 Cathy O' Neil's ORCAA and Joy Buolomwini's Algorithmic Justice League take a more technical approach, and initiatives such as Cansu Canca's AI Ethics Lab apply a social science lens to de-biasing AI. See "Governance" play for more information on auditing algorithms and different options.

142 West, S. M., Whittaker, M. & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. AI Now Institute. Retrieved from https://ainowinstitute.org/discriminatingsystems.pdf.

143 Noor, E. (2020). How we can ensure AI develops as a force for good rather than harm. World Economic Forum. Retrieved from https://www.weforum.org/agenda/2020/01/how-we-can-ensure-ai-develops-as-a-force-for-good-rather-than-harm.

144 Sundar Pichai, CEO of Alphabet and Google called for government regulation to complement tech firms' internal guiding principles and rigorous review processes for ethical AI. In his Financial Times opinion piece, he outlined existing regulatory frameworks that could be drawn from, and emphasized the need for the private and public sectors to collaborate in order to successfully navigate issues around using AI for good. Source: https://www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04

145 IBM's Chair and CEO Ginni Rometty told CNBC that AI regulation, carefully crafted to allow for technological

advancement ("precision regulation") is necessary to ensure that innovation can flourish while maintaining a balance with security and ethics. She weighed in on the need to update conventional rules about "values" to apply to technology in this digital age. Source: https://www.cnbc.com/2020/01/22/ibm-ceo-ginni-rometty-calls-for-precision-regulation-on-ai.html

146 Carson, C. (2020, January 9) Personal Interview.

147 Notes from the AI frontier: Tackling bias in AI (and in ... (n.d.). Retrieved from https://www.mckinsey.com/~/media/McKinsey/Featured Insights/Artificial Intelligence/Tackling bias in artificial intelligence and in humans/MGI-Tackling-bias-in-AI-June-2019.ashx

148 This includes: International Bill of Human Rights (Universal Declaration of Human Rights); International Covenant on Civil and Political Rights; and the International Covenant on Economic, Social and Cultural Rights

149 (2018). Human rights in the age of AI. Access Now. https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf.

150 AFOG. https://afog.berkeley.edu/files/2019/07/AFOG_workshop2018_report_all_web.pdf

151 Proxy variable. (n.d.). Retrieved from https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100351624

152 Hao, K. (2018). What is machine learning. MIT Technology Review. Retrieved from https://www.technologyreview.com/s/612437/what-is-machine-learning-we-drew-you-another-flowchart/.

153 Algorithmic Accountability Policy Toolkit (Rep.). (2018, October). Retrieved https://ainowinstitute.org/aap-toolkit.pdf

154 Algorithmic Accountability Policy Toolkit (Rep.). (2018, October). Retrieved https://ainowinstitute.org/aap-toolkit.pdf

155 Hao, K. (2018). What is machine learning. MIT Technology Review. Retrieved from https://www.technologyreview.com/s/612437/what-is-machine-learning-we-drew-you-another-flowchart/.

156 Sciforce. (2020, January 31). Introduction to the White-Box AI: The Concept of Interpretability. Retrieved from https://medium.com/sciforce/introduction-to-the-white-box-ai-the-concept-of-interpretability-5a31e1058611

157 Hulstaert, L. (2019, March 14). Machine learning interpretability techniques. Retrieved from https://towardsdatascience.com/machine-learning-interpretability-techniques-662c723454f3

egal

https://haas.berkeley.edu/equity/