Bias in Emerging Technology

How to make artificial intelligence more equitable

Laurissa Barnes-Roberts • Julia Forrester • Miriam Havelin • Danielle Lim

OCAD U Strategic Foresight and Innovation | Understanding Systems 6011

Table of Contents

Our Team

Laurissa Barnes-Roberts, Julia Forrester, Miriam Havelin, and Danielle Lim are Master of Design graduate students in the Strategic Foresight and Innovation program at OCAD University in Toronto, Ontario. With professional experience in design and marketing in the public, private, and non-profit sectors, we aim to provide a comprehensive understanding of the technologies around us.

Our research interests include digital communications, misinformation, and equity. This service design brief and synthesis map were developed as part of the course Understanding Systems and contributes to the Strategic Innovation Lab, centre for participatory foresight, systemic design and social innovation, at OCAD U (https://slab.ocadu.ca/project/synthesis-maps-gigamaps).

Introduction

Technologies that use artificial intelligence (AI) have become integrated into every part of human life, informing the news people see, the advertisements people are shown, and even the GPS directions people are given. The use of AI is expanding, and the powerful computers and complex algorithms behind these technologies are becoming increasingly advanced as companies rapidly invest in research and development around AI.

Soon, AI will be widely used to help diagnose diseases, drive cars, and police neighbourhoods (Hawkins, 2018; Martin, 2019; Walch, 2019). These uses may seem like futuristic fictions, but they already exist in the world (and are gaining momentum).

What if people learned, however, that diagnostic AI is less accurate for non-male genders (Kaushal et al., 2020)? Or that self-driving cars are less likely to detect pedestrians with darker complexions (Samuel, 2019)? Or that predictive policing is more likely to negatively impact historically marginalized groups, such as BIPOC and LGBTQ+ communities, those living with mental illness, and/or those who are homeless or from low socioeconomic situations (Kenyon, 2020)?

Would people be so quick to accept and adopt these technologies, no questions asked? Or would people treat AI as imperfect and fallible, like the people who create it?

Definitions

In order to discuss the larger system, alignment around the following key definitions is important:

ARTIFICIAL INTELLIGENCE

There are a number of definitions for artificial intelligence and while many of them are similar it is unsurprising that there is no consensus in this field regarding this definition.

The term artificial intelligence was coined by computer scientist, John McCarthy, who defined it as "the science and engineering of making intelligent machines, especially intelligent computer programs" (McCarthy, 2007). This definition focuses on the process and science of developing. Other definitions such as "AI refers to any human-like intelligence exhibited by a computer, robot, or other machine…the ability of a computer or machine to mimic the capabilities of the human mind" (IBM, 2020), are aligned to the characteristics of the completed technology.

There are additional definitions such as the following written for *Britannica* which combine both; "Artificial intelligence (AI), the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.

The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from experience" (Copeland, 2020).

Broad categories of artificial intelligence include weak or narrow AI, strong AI, applied AI, and cognitive simulation (Britannica, n.d; Builtin, n.d; Frankenfield, 2021). AI is further divided and classified in a few other ways based on functionality, capability, field or research techniques.

ALGORITHMS

The definitions for an algorithm are also numerous, but at its most basic an algorithm is a set of instructions to achieve a goal (Downy, 2019; Merriam Webster, n.d). In computer science an algorithm is a set of steps or instructions which allow a computer program to accomplish a task (Cambridge Dictionary, n.d; Khan Academy, n.d).

Artificial intelligence technologies are often programed using a sequence of large and complex algorithms, written by developers and computer scientists. In computer science, algorithms are often referred to as *code* and the process of creating them is referred to as *coding*. The relationship between developers, algorithms, and AI technologies is crucial to understanding bias in this system.

BIAS

A bias is described as a preference or inclination for or against something (Bias, n.d.). Biases are part of what shapes the human experience. Humans use biases for mental efficiency, to aide sorting through information in the world, allowing people to make decisions more quickly (Vinney, 2018).

Because biases are an integral shorthand to human function (Stanborough, 2020), they are not always noticed by the person wielding them. These are unconscious or implicit biases (Hauser, 2018). Some biases are negative stereotypes, and if they are not examined, a person can act in alignment with these biases, even unconsciously, and perpetuate systemic prejudices through discriminatory actions (Buolamwini, 2019).

Definitions

EMERGING TECHNOLOGIES

Emerging technologies, such as AI, are built by human developers. The choices that are made around programing algorithms, curating training data sets, and how the system is validated allow the biases of the developers, the companies for which they work, and the society in which these algorithms work to become embedded into the AI.

MACHINE LEARNING

Machine learning is a form of AI that enables a system to learn from data, identify patterns, and make decisions with having been explicitly programmed to do so (i.e. without additional human intervention) (*Data Science and Machine Learning*, 2020; Machine Learning, 2021). This data can take the form of anything that can be digitally stored – text, images, numbers, link clicks, etc.

Technologies that use machine learning collect as much data as possible about their users so that they can make informed predictions that anticipate those users' future needs (Hao, 2018).

Overview

For over a century now, authors, scientists, mathematicians, and philosophers have been theorizing about machines that could imitate human intelligence (Anyoha, 2017).

The dominant thinking in the 1950s, when the theory began gaining traction, was that human brains and computers were a "species of the same genus" – essentially, they were information processing systems that could "take symbolic information as input, manipulate it according to a set of formal rules, and in so doing… solve problems, formulate judgments, and make decisions" (Crowther-Heyck, 2008; Dick, 2019; Heyck, 2005; Newell & Simon, 1972).

Advances in technology and computing, particularly in the last 30 years, have led to the development of machines with advanced computing and analytical power and the emergence and growth of AI.

Today, AI is integrated into every part of daily life. Voice assistants like Siri and Alexa use AI to interpret what a user is saying and respond to requests. Streaming services like Netflix and Spotify, search engines like Google, and social media apps like Facebook and TikTok use AI to collect data about what content users engage with (including advertisements) and then recommend what users may want to see or hear next (Hao, 2018; Marr, 2019).

Navigation aids like Google Maps use traffic data and historical traffic patterns to recommend routes and predict how long it will take to get to a given destination (Lau, 2020). Even personal banking apps use AI to track typical customer behaviours and flag anomalies to detect fraud (Walch, 2020). While this technology has been unquestionably beneficial, it is not as benign as it might seem.

AI builds assumptions based on the patterns it finds, which allows it to, arguably, "make better decisions than humans because it can take many more factors into account and analyze them in milliseconds" (Gonfalonieri, 2019). Being able to make 'better' decisions than humans does not make AI faultless, though. Because of the humans who design them, algorithms are susceptible to bias, which can become embedded in the technology at several points throughout its lifecycle (Silberg & Manyika, 2019).

Given the fallibility of technologies like AI (and the algorithms upon which they rely), the research conducted for this systems analysis was guided by the following research question:

How might we use a systemic approach to explore the AI ecosystem in order to suggest possible interventions to reduce bias and make the technology more equitable?

Using a systems-based approach to analyze the AI technology lifecycle and ecosystem in which these technologies are embedded, there emerged several possible intervention points. This brief will discuss the scope of the problem, outline the components and major stakeholders/actors, as well as their relationships, and discuss a few of the most influential potential interventions.

This brief is best read in conjunction with the corresponding synthesis map, which visually outlines the contents herein.

Lastly, two case studies have been added to the appendix in order to provide examples of the proposed interventions.

Scope & Boundaries of the System

In exploring the research question, we dissected the levels of systems involved and understand which levels have the most opportunity for intervention. The issue of implicit bias is woven in multiple ways within each level of this system.

Throughout the lifecycle, various activities occur at different levels:

- The **micro layer** describes what is happening on the frontline and behind the scenes and so is both the most visible and least transparent. It is where the AI is first coded by developers and then first introduced to users or customers.
- The **meso layer** moves outward to the local industry actors, as well as the immediate technology ecosystem such as the developers and organizations who create the AI.
- The **exo layer** encompasses a broader ecosystem, which includes government, and the AI technology sector and related industry actors (e.g. health care policy, data security).
- At the **macro level**, the largest societal forces are at work in the background, including societal values and beliefs, and the pressures and demands of capitalism.

A deeper analysis of the micro level was conducted first by looking at the development lifecycle, to explore the different places where bias is embedded into the system and possible leverage points.

The micro level is where the algorithms are coded by developers in technology companies that are competing on speed to market and therefore also where implicit bias is introduced into the technology (Elsbach & Stigliani, 2020).

This was the focus of the lifecycle section of the synthesis map, and the basis for many of the causal loops, although the analysis uncovered possible interventions across the levels of the system from micro to macro.

The process of taking the idea for the technology from a prototype and funding through to launch is referred to as the development lifecycle. In the lifecycle, speed to market and producing a minimum viable product are essential to secure and retain funding for start-up or small enterprises to eventually deploy.

This is an ecosystem which is typically overwhelmingly male, white, and 'techno-heroic' (D'Ignazio & Klein, 2020). Author, game designer, and Georgia Institute of Technology professor Ian Bogost argues that developers "constitute a 'tribe,' separated from the general public…by the exclusive culture of computing education and industry" (2019). These factors combined create a homogeneous environment where different perspectives and assumptions are more likely to go uninterrogated and unchallenged.

As previously mentioned, the applications for AI are vast and extremely varied. In order to explore the system of bias in AI, strategic generalizations have been made. While specific, individual technologies may vary in terms of their lifecycle, company culture, and funding structures, commonalities exist in the ways in which different types of bias are embedded in these technologies.

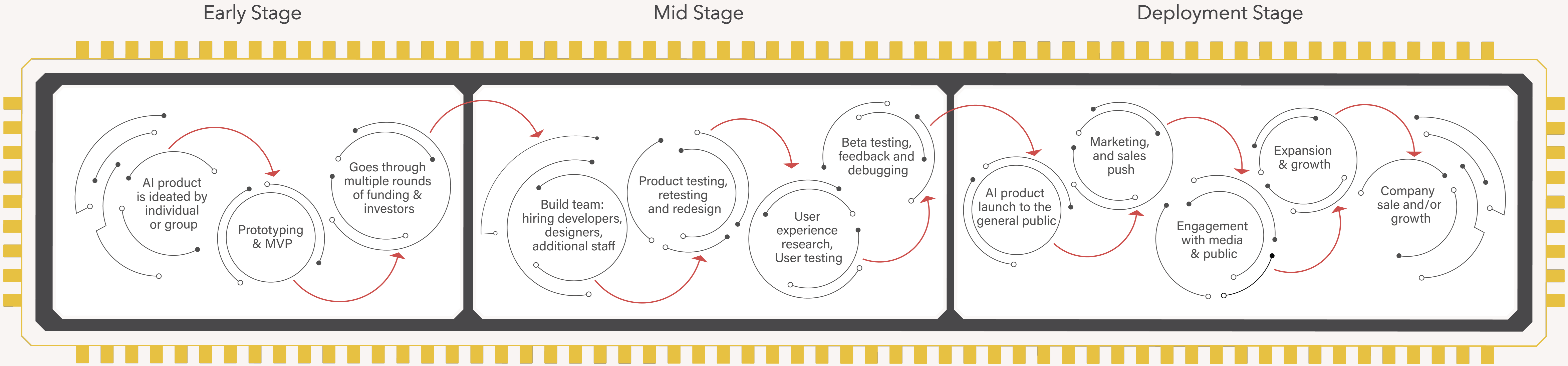
This section will focus on key features of this system and relationships between them and where bias is integrated into them.

Scope & Boundaries of the System

LIFECYCLE OF A PRODUCT

Central to the understanding of this complex system is *the technology*, that is, a given AI product. The applications of AI are nearly limitless, and therefore the specific lifecycles of AI technologies are also extensive. However, in assessing from a system level there are commonalities in the various lifecycles where systemic bias emerges.

Based on research and inference, a generic AI product lifecycle was developed, it is intended to be broadly applicable to multiple AI products. While the specifics of a given AI product lifecycle will likely deviate in some ways from this layout, the broader process and, more importantly, the relationships to other components of the system are intended to be similar across technologies.


Early Stage
Mid Stage
Deployment Stage
AI product is ideated by individual or group
Prototyping & MVP
Goes through multiple rounds of funding & investors
Build team: hiring developers, designers, additional staff
Product testing, retesting and redesign
User experience research, User testing
Beta testing, feedback and debugging
AI product launch to the general public
Marketing, and sales push
Engagement with media & public
Expansion & growth
Company sale and/or growth

This lifecycle allows us to connect components of the system together, along a temporal range.

Scope & Boundaries of the System

The points on the lifecycle fall into three broad categories.

1. **Early stage.** This phase is characterized by ideation and funding. In an established company with internal AI-based technologies, this could be a team pitching a new feature for an existing piece of technology or AI.

   In a start-up, this could be the raison d'être and funding would come from outside investors. In a small company, this could be a business model pivot or a new feature and could be funded internally or require external funding depending on the company's size and assets. Part of this stage regularly involves creating a prototype or minimum viable product in order to pitch it.

2. **Mid stage**. This phase is characterized by team selection, research, further product development, prototyping and rounds of testing, including beta testing, redevelopment and redesign.

   At the end of this phase, a viable product is ready for deployment. In an established company this could look like team member selection, UX research, deployment of a feature/product among a set of users or clients, testing, Q&As and redevelopment. In a small company or start-up this could look like hiring or outsourcing a team, research and narrowing of scope as well as beta testing a product.

3. **Deployment stage**. This phase is characterized by the deployment of a given AI product into general use for the intended market.

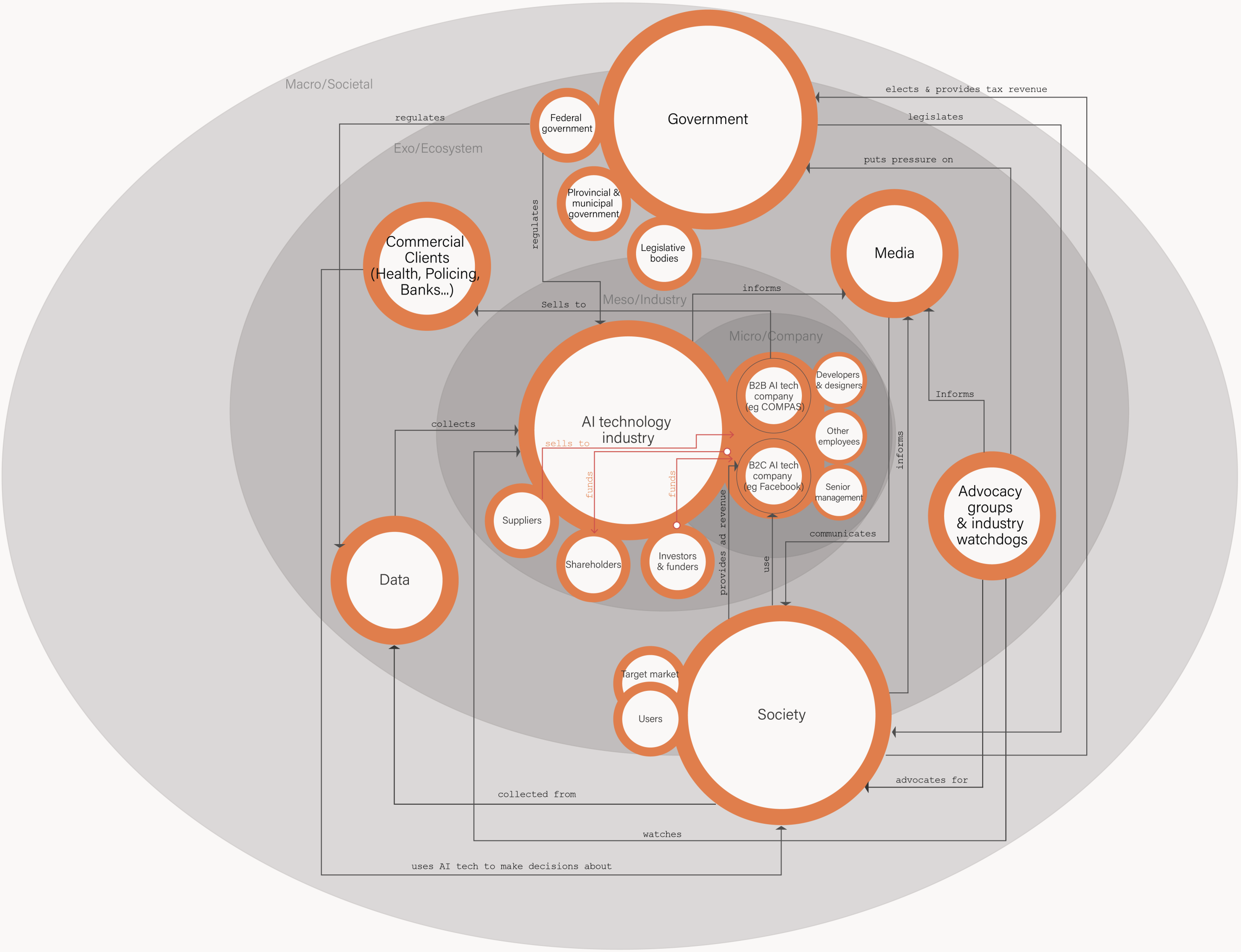
   Depending on the product in question, this phase could look very different among distinct companies, but is broadly characterized by product launch, sales and marketing, and engagement with stakeholders outside of the company. As many technology companies have growth as part of their business models, included in this phase are plans/intention for growth, scaling or expansion.

Scope & Boundaries of the System

MAPPING THE SYSTEM

The system map is an important component of the system, which explores the prominent stakeholders and their interdependencies and influences upon one another.

This system map is not meant to be an exhaustive list of all the stakeholders in each AI system, rather, a high-level overview of major stakeholders more generally.



Scope & Boundaries of the System

The AI industry

Broadly, the AI industry encompasses a number of key actors. Within a given company there are executives and other senior management, developers, designers, and data scientists, as well as other employees. These companies are also engaged with funders, investors as well as shareholders, in some cases.

Generally, these companies can have two main functions; they can be involved in business-to-business (B2B) AI technology or they can be involved in business-to-consumer (B2C) types of technology, (though some large technology companies do both). These distinct characterizations usually change the nature of the AI in use.

Commercial Clients

Business-to-business AI companies supply commercial clients with software or hardware that allows them to accomplish their goals. These types of AI technologies are generally implemented as a tool to accomplish an end purpose.

For example, COMPAS is a technology that assigns a score (between 1 and 10) to a given person who has committed a crime. The score is intended to indicate the likelihood that the individual will reoffend (Angwin et al., 2016).

The technology is used by certain judicial systems in the US to help in deciding sentencing and parole possibilities for offenders. Another example is AI technologies used to pre-screen patients for specific diseases. Although in their early stages, some of these technologies are being implemented through partnerships between hospitals and AI companies (Daley, 2018).

Due to the nature of AI, commercial clients and technology companies regularly work in partnerships when building technology for specific fields such as health, human resources, and policing among others.

Society

Society as a whole is considered one of the stakeholders as there are very few individuals who are exempt from the influence of AI. Whether knowingly or unknowingly, virtually every individual in society has provided data that has been used to train AI algorithms.

User

Within society we have a subdivision of individuals considered *users*. A user is an individual who uses specific technologies or machines (Merriam Webster, n.d). Users provide additional data to technology companies through the devices that they use and products they engage with. The relationship between users and technology companies is layered, while they benefit from the product, they also provide the technology company with revenue in various ways such as data sales, and ad revenue from engagement on their platforms.

Target Market

Another subset within society that is relevant in this system is the *target market,* the group of consumers at whom a specific product is aimed (Kenton, 2021). Technology companies, like any other businesses, have target markets where they feel their technology would be most successful (i.e. generate the most sales) and thus direct their marketing efforts towards these groups. Unlike with commercial clients, technology companies rarely collaborate with users or target markets to develop technologies in conjunction with their needs.

Scope & Boundaries of the System

Government

The different levels of government legislate and regulate many parts of this system including: the technology and AI industry (to an extent), data usage and collection, and lastly society in general. Governments are also recipients and users of AI technologies; many government bodies work with technology companies to integrate artificial intelligence into government policies, practices, and endeavors. Governments are heavily influenced by society, through societal pressures and voting.

Media

The media is a key player in this system. The media communicates to and with society by broadcasting newsworthy content to the public and sometimes being informed by whistleblowers and informants which can have huge consequences.

For example, Google came under scrutiny recently when a secretive contract with the US Department of Defense was revealed in the media. The ensuing firestorm of criticism from within and outside the company caused them to not renew their contract once it ended (Vox, 2021).

Data

Data is a prominent stakeholder in the system. Data is what AI algorithms are trained on and therefore directly impact the final product.

Data is collected from society; it is a commodity that can be bought and sold. Thanks to the advancement of technology, data collection, use and storage is occurring at an unprecedented rate. AI technologies use existing data to train their technologies and collect data on their users in order to have more material to train their algorithms with.

The interactions between these various stakeholders as well as the company culture and the culture of the AI and technology industry, are what cause bias to become embedded in systems of AI.

Scope & Boundaries of the System

In visualizing the lifecycle of a piece of AI technology, there are multiple places in which bias can inadvertently be brought into AI. Through research, four key places were identified where bias is brought into the system. These correspond to the selected intervention strategies within the system, where bias in AI is prominent:

1. The first place is funding. "Investors are inherently biased, and intuition alone cannot consistently drive good decisions" (Bueschen, 2015). Investors have many unconscious biases that impact the way they fund potential technologies including similarity bias, local bias, anchoring bias, and gender bias (Bueschen, 2015).

2. The next stage is hiring and company culture. Biases in hiring practices have long been documented. These biases can exhibit themselves as gender bias, racial bias, and ageism. The culture of an organization or even an industry can have an impact on potential biases built into their technology. There are many levels of culture within the AI industry–individual technology companies have their own culture; the industry as a whole has a culture–and this culture comes with its own embedded biases at different levels.

3. The third place where we see bias being built into the system is through data collection. As mentioned previously, data is what's used to train an AI algorithm, and so when the data pool is biased, the outcome will be biased.

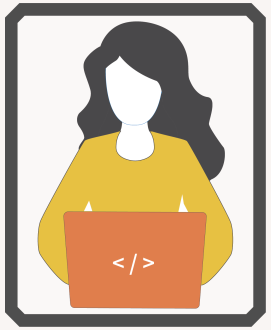
4. Lastly, there is a bias in public perception of technology and technology companies. Oftentimes AI and technology is perceived as being neutral and without bias. This public perception leads to the unquestioning use and deployment of AI technologies in arenas of public and commercial life without proper scrutiny or checks and balances.

These identified intervention points, which will be explored in more depth in subsequent sections, correspond with the crucial points of bias which emerge from the exploration of this system, its stakeholders, and their interactions.

Stakeholders & Actors

The relationships in the AI system are scattered, where some play more active roles, and some are acted upon in the system.

AI Developers



The AI developers have an advantageous position. They closest to the product with full access to the proprietary AI. They can make direct changes in how the AI operates and functions. They have their own organizational goals to achieve and the desire of freedom to operate and innovate.

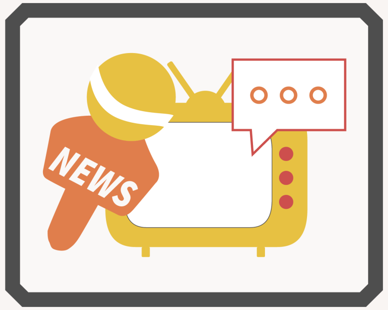
Government

The government has a similar level of power as the developers; however, they are not as agile as the pace of innovation outpaces regulation. While AI is rolled out publicly, governments struggle with gaining full access to AI information, lack technical capacity, and are required to make grounded policies to update or introduce new regulations. As such, regulation in this sector is lagging (Mozilla Foundation, 2020). Still regulatory and financial bodies have had the most pronounced impact on shaping the cycle of innovation (Henton & Held, 2013) and AI development.

Investors (VCs, angel investors)

Investors play a critical role as they control a large portion of financing of AI technologies and innovations. As the industry is driven by commercial incentives and growth, their needs often are prioritized by AI developers and companies.

Media and Shareholders



Both these groups fall into the same category, especially as it relates to the proprietary algorithms and technology. They are potential leverage points, but similar to the government have less access to internal information.

Civil Society Groups

These groups can work to help counteract the lack of information by helping to disseminate information, advocate, and create more awareness. However, they currently lack strong platforms to raise issues of concern.

Stakeholders & Actors

General Public

The general public may be evaluated by the AI. They may be applying for a bank loan, job, parole, or looking for recommended products. AI and machine learning are ubiquitous and often appear in high-risk situations, but their impact and pervasiveness may not be apparent to the general public. They may also be unaware of the lack of transparency around how their information is used. How AI works is not transparent (Buolamwini, 2019).

The general public has low knowledge and power yet can be the most impacted by the technology. Decisions are regularly made about them without their consent or awareness. Some of these AI assessments may not be accurate. Numerous incidents have been documented that show that AI can reinforce challenges faced by individuals who are marginalized. Public demands for transparency, protections to prevent corporate abuse, and data security have been rising, but overall general awareness is still relatively low.

Interventions

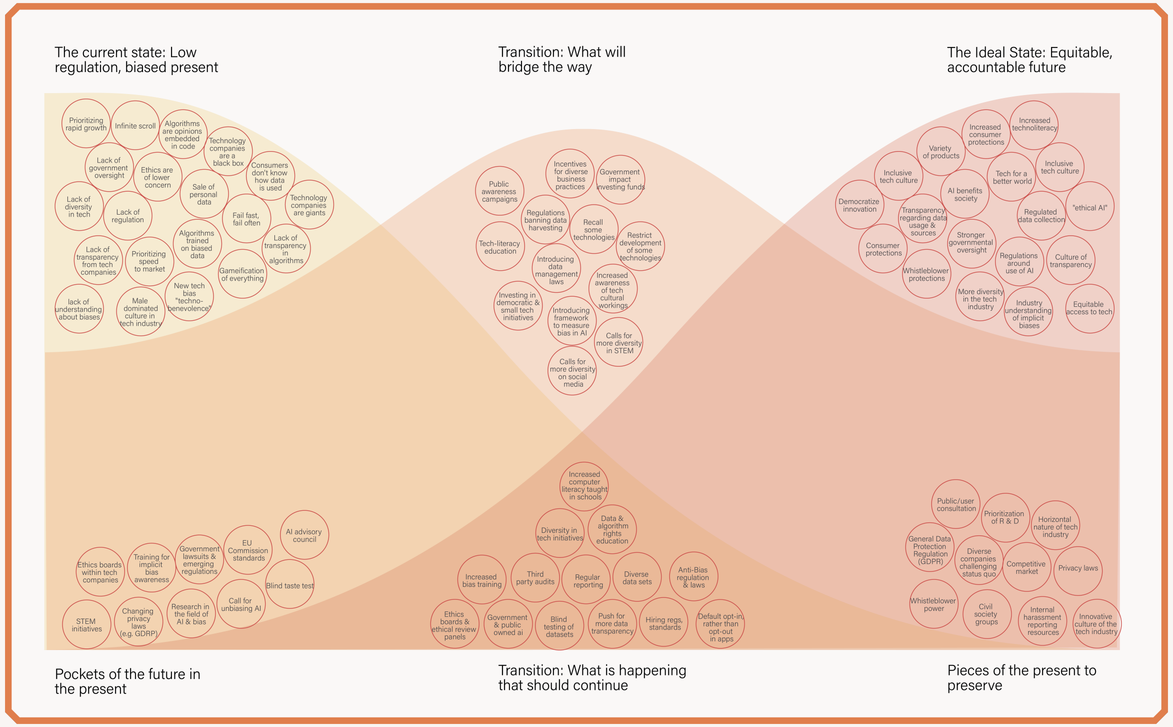
At four key stages of the AI development lifecycle, where bias may most be embedded in the final product, intervention strategies have been identified.

Each intervention strategy has a combination of leverage points, influenced by Donella Meadow's *Leverage points: Places to Intervene in a System* (1999), and tactics targeting the relationships between stakeholders or opportunities to restructure elements of the system.

These recommendations, if implemented, would have a resounding impact on the system.

The comparison of the current state to the ideal future shows a shift in power, checks and balances, and an environment more conducive to equity in the future. To reach this more equitable and ideal state will take cooperative effort, time, and investment as well as implementation of multiple intervention strategies for different levels of the system.

The intent of this brief is to encourage, and we hope policymakers can begin to incorporate areas of these interventions. Private AI developers, likewise, can consider the recommendations and make proactive strides towards equitable AI.


The current state: Low regulation, biased present
Transition: What will bridge the way
The Ideal State: Equitable, accountable future
Prioritizing rapid growth
Infinite scroll
Algorithms are opinions embedded in code
Technology companies are a black box
Lack of government oversight
Ethics are of lower concern
Consumers don't know how data is used
Sale of personal data
Technology companies are giants
Lack of diversity in tech
Lack of regulation
Fail fast, fail often
Algorithms trained on biased data
Lack of transparency in algorithms
Lack of transparency from tech companies
Prioritizing speed to market
Gameification of everything
New tech bias "techno-benevolence"
lack of understanding about biases
Male dominated culture in tech industry
Public awareness campaigns
Incentives for diverse business practices
Government impact investing funds
Regulations banning data harvesting
Recall some technologies
Tech-literacy education
Restrict development of some technologies
Introducing data management laws
Increased awareness of tech cultural workings
Investing in democratic & small tech initiatives
Introducing framework to measure bias in AI
Calls for more diversity in STEM
Calls for more diversity on social media
Increased technoliteracy
Increased consumer protections
Variety of products
Inclusive tech culture
Tech for a better world
Inclusive tech culture
Democratize innovation
AI benefits society
"ethical AI"
Transparency regarding data usage & sources
Regulated data collection
Consumer protections
Stronger governmental oversight
Regulations around use of AI
Culture of transparency
Whistleblower protections
More diversity in the tech industry
Industry understanding of implicit biases
Equitable access to tech
AI advisory council
EU Commission standards
Government lawsuits & emerging regulations
Training for implicit bias awareness
Ethics boards within tech companies
Blind taste test
Call for unbiasing AI
Research in the field of AI & bias
Changing privacy laws (e.g. GDRP)
STEM initiatives
Increased computer literacy taught in schools
Diversity in tech initiatives
Data & algorithm rights education
Increased bias training
Third party audits
Regular reporting
Diverse data sets
Anti-Bias regulation & laws
Ethics boards & ethical review panels
Government & public owned ai
Blind testing of datasets
Push for more data transparency
Hiring regs, standards
Default opt-in, rather than opt-out in apps
Public/user consultation
Prioritization of R & D
Horizontal nature of tech industry
General Data Protection Regulation (GDPR)
Diverse companies challenging status quo
Competitive market
Privacy laws
Whistleblower power
Civil society groups
Internal harassment reporting resources
Innovative culture of the tech industry
Pockets of the future in the present
Transition: What is happening that should continue
Pieces of the present to preserve

Interventions

INTERVENTION 1: EQUITABLE GOVERNMENT-SUPPORTED FUNDING

Layer: Exo (Ecosystem)

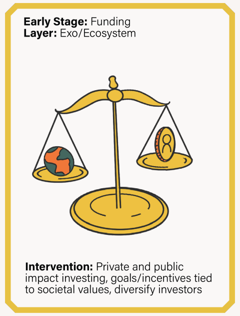
Lifecycle stage: Early Stage

Key actors: Government, investors, start-ups

Leverage points: Government oversight, public funding, extension of regulations, supporting current impact investing infrastructure

Timeline: 5-10 years. Government needs to set up impact investment funds and incentives and joint partnerships. Investors see impact of new funding after five years, as companies start to see returns (Tepper, 2020).

Support Processes: Government investment funds, government partnerships and AI procurement process, government sets impact incentives to achieve societal goals, investment industry transparency requirements.



Early Stage: Funding
Layer: Exo/Ecosystem

Intervention: Private and public impact investing, goals/incentives tied to societal values, diversify investors

Context

Funding for start-ups is a critical step for its growth and longevity. One third of the 'Innovation Success Triangle' by Robert D. Atkinson, is the business environment. The other two are political and social institutions.

The US is the leader in the private venture capital (VC) industry and has been growing in recent years, however, government funding and development of federal innovation systems are lagging (Atkinson, 2014). While some organizations may grow organically, funding during the early stage of the AI production lifecycle, can rapidly scale the organization, and dramatically alter the success and trajectory of the business.

In the influence map, investors are a key and powerful input into AI development.

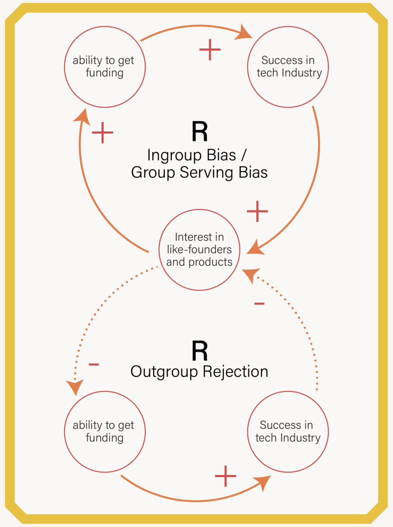
Funding, through angel investors and VCs, decides who gets to make technologies and human biases and values are embedded throughout the process. Valuation is highly dependent on the founder, initial team (Chae, 2019), and prototype (as seen in the product lifecycle as key activities). Those who have previous start-up experience, received funding before, or have an existing relationship with the VC firm are at an advantage (Harroch & Sullivan, 2018), perpetuating the system archetype, 'Success to the Successful' (Braun, 2002).

In this archetype, there are two reinforcing loops. While those who fit the mold of what investors look for and who have had previous success gain more funding, entrepreneurs who do not fit their norms continue to receive less funding. The system is controlled, rigid, and closed, based on key firms, personal relationships, and similar values and beliefs.

Interventions

VCs select and finance companies, as well as determine the direction of knowledge, learning, and research (Ferrary & Granovetter, 2009). They have direct influence on how start-ups operate in terms of financing, strategic assistance, reputation, access to networks (partners, resources, and staffing), and governance structure. They may also stipulate veto rights of certain business decisions (Harroch & Sullivan, 2018).

Many conflicts arise between entrepreneurs and investors due to asymmetry of information and power. They may differ on risk, objectives, management style, and values. This can ultimately manipulate the start-up or alter the innovation drastically (Fassin & Drover, 2017).



Looking closer into the individuals that make up the VC and funding industry, there is a glaring lack of diversity. White employees make up 72% of the VC workforce. Asian and Pacific Islanders make up 18%, Hispanics, 7%, and Black employees 4%. The workforce is also 77% male (NVCA, Venture Forward, & Deloitte, 2021). Additionally, many Black venture capitalists do not make it past an 'associate' or 'principal' title. Silicon Valley's most high-profile VC firm, Sequoia Capital, has no Black partner.

Gender bias also has an evident impact (Bueschen, 2015). 97% of VC-backed ventures have male CEOs (NVCA, Venture Forward, & Deloitte, 2021).

This labour force does not experience much change as the turnover of senior employees is low and hiring has often been based on individual networks (NVCA, Venture Forward, & Deloitte, 2021).

VC firms tend to prioritize founders with similar backgrounds as them, such as in education and work experience. They often select ventures that are geographically close to their firm. Historically, 50% of firms have been with in 375 km (233 miles) of the VC office (Bueschen, 2015). These biases and the networks of funding that have been traditionally set up, perpetuate the biases in the technologies they invest in.

The vast amount of funding seen in regions like Silicon Valley have directly contributed to numerous, impressive innovations and technologies, but may not align with societal values. The current funding environment creates a gap in AI innovations that may prioritize goals other than profit or do not show financial potential early on. Investors may put pressure on companies to release AI products to meet market demand, increase profits, and reduce financial risks. This reduces the time to validate the algorithms, which opens the door for unintended consequences, such as perpetuating biases.

Interventions

Strategy Outline

The intervention needed at this stage of product development is impact investing from private and public entities and diversification of the funding environment to reduce bias.

Impact and Ethical Investing

Impact and socially responsible investing (SRI) have a part to play in changing how technology is created. Impact investing involves investments that would ultimately benefit society. SRI is selectively making investments based on social values and ethics. Environmental, social, and governance (ESG) may also be incorporated in evaluating an investment (Zhou, 2019).

Responsible investing is starting to make up a large portion of total investing. In 2018, 25% of managed assets in the US were impact investing. The current landscape has yet to mature, but has shown early success in creating more ethical and diverse products, as seen with green tech. The rise of B Corps and younger, more socially conscious investors have contributed to that growth (Mozilla Foundation, 2020). Steps towards a more equitable investing ecosystem include measurements, incentives, and goals in line with social values, private and public impact investment funds, and diverse investors.

Implementation

Government Investing

There are trade-offs and challenges with impact investing. Administrative and opportunity costs reduce earnings and discourage VC firms. Impact investing is not rewarded by the system.

Investors face fewer disincentives when there are negative consequences compared to large bonuses tied to financial gains. Nassim N. Taleb and Constantine Sandis raise this issue in *The Skin In The Game Heuristic for Protection Against Tail Events*. Investors should be more responsible for ups and downs in the market and the reverberating impacts they have on society. Currently when there are negative impacts, the blame and onus is absorbed by others (Taleb & Sandis, 2014).

This is where government intervention could create the right incentive structures, investment rules, and market infrastructures (Koenig, 2016) in line with their innovation and economic mandates. By accounting for externalities, value could be more accurately captured in financial markets, and they could better regulate investors (Schwartz & Finighan, 2020).

Direct government involvement, oversight, and investing can dramatically alter the financing and innovation space. It can target the demand (by building the capacity of impact investors, public procurement) and supply (by public matching and investment) sides of social impact, as noted in the Policy Framework from PCV InSight and The Initiative for Responsible Investment.

Several governments are already taking control of technology funding. The German government pledged €3 billion for AI R&D and Europe, stated to increase their investments to over €20 billion (Mozilla Foundation, 2020).

Interventions

The European Commission, which is highly regarded as one of the leading policies in ethical AI, is developing investment funds specifically for AI. Another proposed initiative by the European Commission is the creation of public-private partnerships. This will strengthen collaboration and alignment on goals (European Commission, 2020).

Other ways to increase supply include pay-for-success commissioning, capacity-building grants, and reducing barriers to financing (PCV InSight & The Initiative for Responsible Investment, 2014). The US government is not highly involved in the AI technology industry. Their participation can help direct innovation and diversify economic growth. The government's own procurement, use of AI, incentives, involvement as a stakeholder can energize the market for ethical AI (Mozilla Foundation, 2020).

Governments can steward the market through awareness, supporting intermediaries, reducing barriers to entry, and reducing transaction costs (PCV InSight & The Initiative for Responsible Investment, 2014). Fundr is an example of a new structure that creates a pathway for diverse founders to enter the industry and scale up (*Fundr*, n.d.). It sets new standards on how to evaluate, manage, and discover companies. Their AI selection process is designed for fair evaluation, to remove bias from investor searching, weighs underrepresented founders more favourably, and creates diversified portfolios (Tepper, 2020). Start-ups can then find investors that align with their values and be included in the pool of potential investment recipients.

Dedicated government budgets and funding structures can help balance the current flows of investment

Measuring Impact

Impact investing requires a paradigm shift; a company can still be profitable and help solve world problems. Impact investors are better designed to take on more risks and they should be rewarded, compared to traditional VCs who are more risk-averse and, so, target a homogenous set of start-ups. Actionable internal mechanisms, such as changing measurements to consider risk, reward, and impact, are a starting point (PCV InSight & The Initiative for Responsible Investment, 2014).

Investors need to set social and environmental impact metrics that are tied to financial incentives in the short and long-term. This introduces key values and ethics at the forefront (Global Impact Investing Network, 2011) of daily activities, exposes effects on the wider society, and upholds fiduciary duties.

For example, BlackRock requires companies to show positive impact before receiving disbursements, and Vox Capital has incorporated impact achievements in employee reviews (Global Impact Investing Network, 2018). Governments can also use this in their funding and private-public partnerships. These measurements can begin to reorient investor and start-up behaviour.

Diversifying Venture Capital

Diversity in the industry equally needs to be addressed to further democratize who holds power and capital, and support impact investing ideals.

Interventions

Goldman Sachs reported that all-female or mixed-gendered investment teams outperformed all-male teams. Similar studies have been conducted with people of different ethnic backgrounds and found a positive correlation between diversity and performance, number of acquisitions, and IPOs. Diversity is also reinforcing, as a diverse team will attract more people from different backgrounds (NVCA, Venture Forward, & Deloitte, 2021). Diversifying the industry requires challenging dominant beliefs and values. A diverse ecosystem of investors and portfolios needs to be created.

To achieve this, investment firms need to look internally to make hiring changes. Governments and the industry can advocate and incentivize for diverse teams and enforce accountability and transparency by including diversity in their funding and partnership criteria.

Limitations

Capturing and measuring externalities and accurately attributing it to a source is not perfect. The impact of investments is not realized for at least five years, as companies start to see returns, creating a delay (Tepper, 2020). This may cloud how incentives and regulations are formed, and in turn, impact results of investing. Input from a wide range of experts, such as AI ethics boards and the entrepreneurial community, can inform setting targets. For example, investors may work with partners to refine start-up selection criteria. Impact investing is more effective with the cooperation of private investors and must be able to self-organize without the dependence on government.

Change is needed in the funding systems for AI technologies and start-ups. Financial structures need to innovate, this can be achieved through government involvement in the market; creating measurements tied to societal impact, advocacy, supporting impact funds, and the creation of new structures that support diverse businesses and investors. VCs and investors need to operate in a way that is both beneficial for them and for society. The inclusion of the wider society stakeholders strengthens the power of underrepresented voices. The involvement of the government as a policy maker or an investor will allow them to better align their policies on innovation and AI.

Interventions

INTERVENTION 2: DIVERSE HIRING AND COMPANY CULTURE

Layer: Micro/Company

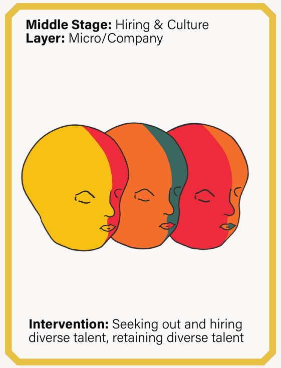
Stage: Middle

Key Actors: AI developers, technology workers and executives

Leverage points: Current recruiting and screening processes, technology executives

Timeline: 2-10 years. Talent scouting and screening practices can be changed within months to bring more diverse talent on board, but it will take a couple years to examine whether the company has retained these diverse hires. The greatest contributor to retention is inclusion, which will require a cultural shift within the industry over the span of many years.

Support Processes: Technology industry leaders, universities



Middle Stage: Hiring & Culture
Layer: Micro/Company

Intervention: Seeking out and hiring diverse talent, retaining diverse talent

Context

Silicon Valley has become a global hotspot for technologic innovation, with companies like Apple, Google, Facebook, and Tesla calling it home (Weinberger et al., 2020). Being located so closely to Stanford University and University of California, Berkeley, these technology companies have formed a relationship with the schools nearby, gaining research and talent from each, while in turn generating billions in economic benefits (Huffman & Quigley, 2002).

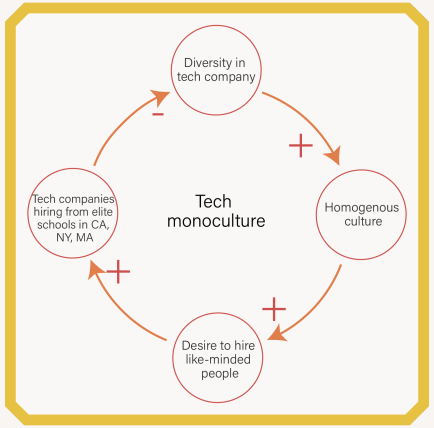
However, this relationship has also led to a level of homogeneity amongst technology companies in Silicon Valley. Representatives of a given firm are "more likely to hire someone from their alma mater, all else being equal" (Huffman & Quigley, 2002). That means that if companies from Silicon Valley are hiring many of their employees from nearby universities, those employees are more likely to hire future employees from those same schools. Given that the majority of students at Stanford and UC Berkeley are white or Asian in ethnicity (*Stanford University*, n.d.; *University of California-Berkeley*, n.d.), those demographics are then reflected in the region's technology companies. Technology employees are also overwhelmingly men (Harrison, 2019).

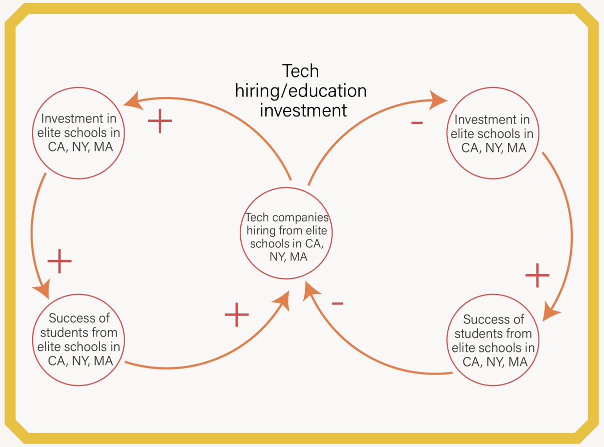
Technology firms do not hire exclusively from Stanford and UC Berkeley, but they do tend to stick to schools in the "traditional tech clusters of California, New York, and Massachusetts" (Chakravorti, 2020).

Therefore, not only are technology company employees largely educated at the same institutions, they are also highly similar demographically. As such, they likely approach their work with a similar worldview and possess the same entrenched biases, which are reflected in their work (whether implicitly or explicitly).

Interventions

The lack of diversity in technology has long been documented, and although many large technology companies have made commitments to diversifying their workforces, little progress has been made. This is especially true when it comes to the backbones of these companies, technical workers – developers, engineers, and data scientists – who research, design, and build the technology and AI products (Harrison, 2019). Even more so than technical professionals, technology leadership lacks diversity, with white people representing 83% of executives in the industry (DeNisco Rayome, 2018).


Diversity in tech company
-
+
Tech companies hiring from elite schools in CA, NY, MA
Tech monoculture
Homogenous culture
+
+
Desire to hire like-minded people


Tech hiring/education investment
Investment in elite schools in CA, NY, MA
+
-
Investment in elite schools in CA, NY, MA
+
Tech companies hiring from elite schools in CA, NY, MA
+
+
-
Success of students from elite schools in CA, NY, MA
Success of students from elite schools in CA, NY, MA

Strategy outline

To intervene at this stage of the AI product lifecycle, efforts should be focused on hiring and retaining diverse talent.

Hiring

In many ways, the biases embedded in human-made AI reflects those of humans who created it. When teams lack diversity, issues that affect those in marginalized groups may not be top of mind. It is for this precise reason that facial recognition AI, for example, developed by all-white teams, consistently fails to properly identify and/or distinguish between non-white faces (Li, 2020). If these teams had included even one person of colour, ensuring the AI could operate effectively across skin tones would likely have been a much higher priority. This ability to challenge entrenched biases is one of the reasons why having a diversified workforce is so key.

Retaining

It is not enough to find diverse talent though; companies must also work to retain them.

Technology firms have been shown to have much lower retention rates of Black and Latinx employees than their white counterparts (Harrison, 2019). Charles Isbell, Dean of Computing at Georgia Tech, argues that diversity means nothing without integration. “The integration of women, people of color, and other underrepresented voices would mean that the behavior of the entire industry would change as a result of their presence in that community,” says Isbell (Bogost, 2019). For that to happen, entrants to the technology field need to be “capable and confident,” but the industry also needs to be willing to include and accept them (Bogost, 2019).

Interventions

Thus, a cultural shift towards diversity and inclusion is required for real, meaningful change to take place.

Implementation

Hiring

Recruit Outside “Traditional” Educational Backgrounds/Skillsets

A proposed solution for technology companies to diversify their workforces is finding diverse talent where they are – in states like Georgia, Texas, Delaware, Virginia, Connecticut, and Maryland, which have more affordable costs of living and higher proportions of minority groups with STEM degrees – rather than expecting diverse hires to come to them (Chakravorti, 2020).

However, hiring candidates from unconventional educational backgrounds (including those without degrees) or skillsets, especially for positions which have a shortage of traditional candidates, has also been shown to increase the diversity of teams (Rosenbaum, 2019).

Objective Interviewing

Once diverse talent is identified, it is important that they are then given a fair chance at being hired. Research has shown that human and AI resume screening alike are inherently biased, with candidates being discriminated against based on “race, religion, national origin, sex, sexual orientation, and age" (Li, 2020). To mitigate this bias, hiring managers should use more objective interview techniques, such as project-based assessments. These assessments have candidates demonstrate their abilities, rather than just speaking to them on a resume, and would give candidates with ‘unconventional’ education or skillsets an opportunity to join the technology field (Li, 2020).

Retaining

Bias Training

Bias on its own is not inherently bad. However, it is important that those working within the technology industry, from technical workers to senior management, are taught to be both aware of their biases and understand how those biases impact their mindset, behaviour, and work (Woo et al., 2018). Without this cognizance, the biases of those in technology are embedded into their organizations and the products they create and are often only noticed when they fail to work for marginalized groups who were not included in the development process.

Leading by Example

For diversity and inclusion to become more engrained into the industry’s culture, there must be support from senior management, who can demonstrate steadfast commitment to diversity efforts and challenge reluctance from lower-level leaders (Dickey, 2019). Articulating and modeling their expectations of inclusion sets the tone for the rest of the company and lets diverse employees feel more comfortable being their authentic selves in the work environment (Gassam Asare, 2018; Woo et al., 2018).

Workers Organizing

While a top-down approach to equity, diversity, and inclusion are invaluable, it must also be possible for workers to drive change from the bottom up if they feel leadership is not doing enough. Some technology workers have begun organizing in response to frustrations with their organizations' corporate cultures, such as when 20,000 Google employees and contractors walked off the job "to protest the company's handling of sexual harassment allegations" in November 2018, just months after a smaller group of 3,000 employees protested Project Maven, Google's military contract with the Pentagon (Fernández Campbell, 2019).

These protests were met with retaliation from Google, who punished organizers of the November walkout. Leslie Miley, a former engineering manager at Twitter, Google and Apple, argues that organizing is necessary for workers to affect change, and that doing so may require unionizing, since those benefitting from the system (i.e., executives) have no incentive to change (Dickey, 2019).

Mentorship

Mentoring can be a great retention tool for companies. Focusing mentorship efforts on employees from underrepresented and/or marginalized groups (e.g., women, BIPOC employees) especially, can result in a more engaged, diverse workforce (Woo et al., 2018).

Mentorship has also been shown to help marginalized employees reach upper management positions (*Mentoring*, 2011), which is critical given the lack of diversity across top executives in the technology industry.

Employee Resource Groups

Employee Resource Groups (ERGs), or affinity groups, are "voluntary, employee-led groups that foster a diverse, inclusive workplace aligned with organizational mission, values, goals, business practices, and objectives" (*Employee Resource Groups*, 2021).

These groups help new employees learn about their company without being overwhelmed with new information, and the equity and sense of belonging they foster has been demonstrated to positively impact retention rates (Gassam Asare, 2018; Woo et al., 2018).

Incorporating EDI as Start-ups

While adopting the above hiring and retention practices is crucial to the diversification of the technology industry, it is worth noting that it is more difficult for companies to make meaningful change after reaching a certain threshold of employees. As such, it is paramount that companies incorporate equity, diversity, and inclusion (EDI) efforts as start-ups to challenge the dominant technology monoculture (Dickey, 2019).

Measuring Impact

The impact of the above hiring and retention recommendations can be measured through diversity reports, which have been generated and released to the public by companies like Apple, Facebook, Google, and Microsoft since 2014 (Harrison, 2019). These reports include information around hiring, retention, and representation (including at the leadership level) by gender and ethnicity. Significant increases in the hiring and retention of women, Black, Latinx, and Indigenous would show meaningful impacts of the actions outlined above on the diversity of the industry.

While these reports provide valuable metrics around the diversity of these companies, they are also produced voluntarily. Were there a technology oversight organization, diversity reports could be standardized and mandated across the industry to ensure the impact of diversification efforts could be measured equitably.

Limitations

While there is currently an economic argument for increasing diversity in the technology industry, there will be no real change until there is a shift towards morally motivated diversity efforts according to Charles Isbell (Bogost, 2019). The lack of change in the composition of technology firms in recent years, despite these companies making public commitments to diversifying their workforces, is reflective of this.

However, Amy Webb, a professor at New York University and the author of *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity*, argues that there is a deeper problem than the lack of representation of minority groups like women and BIPOC folks in technology: the industry environment itself. "Scale, market share, and speed are the top priorities in the industry, driven by the fierce competition between technology firms who share little but the exclusive culture of computing education and industry" (Bogost, 2019).

So long as this culture and competition exist in the technology industry, so too does a divide between technology creators and users, which may be much more substantial than any demographic or educational divides between technology employees.

Interventions

INTERVENTION 3: ETHICS REVIEW ASSOCIATION

Layer: Meso/Industry

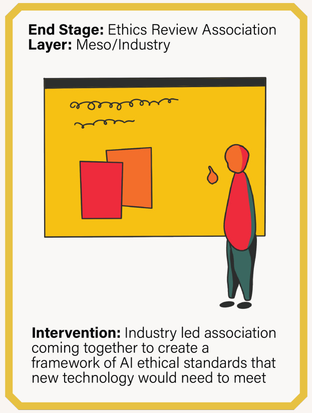
Stage: End/Deployment

Key actors: Executives, industry experts, government

Timeline: 1-5 years. Evidence for this type of intervention is already emerging in the industry. Once in the works, the estimated timeline would be between one and five years depending on the nature of the association and its geographic reach.

Leverage points: Current industry experts, similar industry practices, governmental policies

Support Processes: Government, Intergovernmental organizations, public policy



End Stage: Ethics Review Association
Layer: Meso/Industry

Intervention: Industry led association coming together to create a framework of AI ethical standards that new technology would need to meet

Context

While new developments in AI and its use emerge daily, there is very little oversight internally or externally to regulate these new technologies and assess their impact. Societal and media pressures at times force companies to evaluate the ethics of their own technology however, this is largely an anomaly and not the norm.

The driving goals for large and small technology companies is to generate revenue, be that through sales of their technology or through growth of users on their platforms. This drive can sometimes lead to practices that are ethically ambiguous or the use of AI that has biased outcomes and further entrenches social disparities. This could trigger a company to think about their technology development and set about improvements and changes, or a company could choose to ignore these ethical questions in favour of growth and revenue.

Currently, there is no way to know whether or how often these ethical decisions happen within a given company. It is generally not incumbent upon the company to disclose this information and there are currently no enforcement requirements forcing companies to disclose internal discussions around AI and algorithmic decisions. There have been high profile instances of companies being held to account for a technology or development which is proven to be unethical or result in biased outcomes, but usually the technologies created are trademarked intellectual property that is not shared with the public.

The crux of the issue is that it is entirely up to each company to make these ethical decisions about their technology independent of external scrutiny or involvement.

Interventions

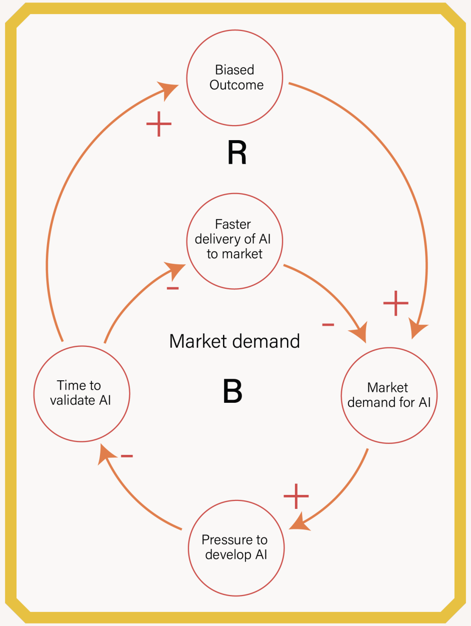
These companies then deploy and/or sell their technologies without needing to disclose any potential biased outcomes, which could result from their use. This opacity and privacy around developed technologies is intended to protect companies from having their IP copied or mimicked by competitors. However, the use of large amounts of personal data from individuals, as well as the deployment of many of these technologies in the wider public, suggests that some level of transparency of algorithms, and the accuracy and level of bias of the outcomes is important for the public to know.

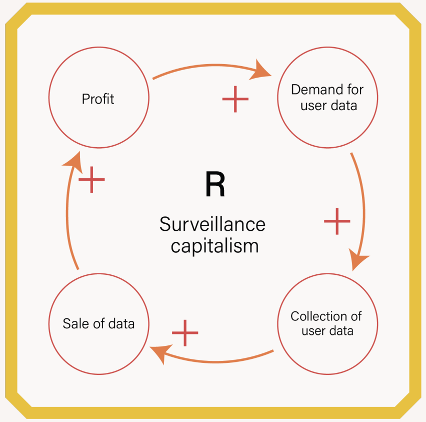
Employees at companies like Google, who have very horizontal working cultures, have been instrumental in voicing concerns over unethical AI practices (Ghaffary & Kantrowitz, 2021). Whereas, at companies like Facebook, decisions about ethics and the pursuit of specific technology are largely made by leadership alone with few exceptions (Hao, 2021).

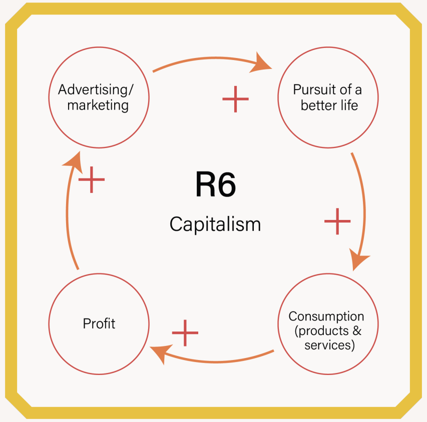
Either way, the companies themselves decide when and how to implement ethical considerations in their technology. It is worth noting that in both companies attempts at internal ethics panels have been more about paying lip service to 'ethics' and less about implementing ethical practices into their technology development (Johnson & Lichfield, 2019; Hao, 2021).

Strategy outline

This intervention strategy suggests the creation of an industry-led association, which reviews emerging AI technology and AI products from new and existing companies. The association would consist of experts and leaders in the industry as well as ethicists and researchers, to create a framework of ethical AI standards that new technology would need to meet in order to be implemented in the wider public.







Interventions

New and emerging technology would be subject to regular review and enquiry by the association. The minimum necessary standards to be met for each technology would vary by industry and type, which is why this intervention model requires industry leaders to be implemented appropriately.

Industry incentives

There are several incentives for the industry to implement an ethics association. First, this would provide much needed industry oversight from individuals who are knowledgeable in the industry. Furthermore, there is a need for an ethical and bias-reducing framework in the industry. As time passes, different governments may feel pressured to step in and this could lead to oversight that is too broad, too restrictive, or ineffectual.

Lastly, there have been several high-profile cases of AI technologies being verifiably biased, often to the detriment of already marginalized populations. These instances may cause people to lose trust in the technology and the companies that create them. With an ethics association, some of this loss of trust can be mitigated.

Implementation

Here, government could play a role in fostering the development of an ethics association. The government's capacity to organize and their funding power could be leveraged to help develop this organization. The government could also provide a base set of regulations/requirements for bias reduction and ethics in AI technology based on government policies. With the potential for long-term involvement in the form of mutual advisory committees and/or government representation on the board or within the organization.

Alternatively, this association could be developed through international collaboration and cooperation, mediated through an intergovernmental association such as the Organization for Economic Cooperation and Development (OECD), which has an AI policy observatory as well as an AI advisory council made up of member states (OECD, n.d). The OECD has the benefit of mediating international issues which require collaboration across distinct stakeholders, such as international tax policy regulations, and could leverage this experience to AI technologies.

A third option would be a group of industry leaders from various companies coming together to create an association. To do this, the company leaders would likely need to involve mediators, and multiple stakeholder groups outside of the industry. A good example of this type of organization is Ad Standards, which is an industry led association regulating advertising in Canada, it provides research, guidelines, receives and reviews public complaints and holds advertisers accountable. The board is made up of senior executives in the Canadian advertising industry (Ad Standards Canada, 2021).

Measuring Impact

The association would have the explicit goal of endeavouring to make AI technology more ethical and bias-free for the public. Measuring the impact of this type of association will be difficult for a few reasons, first, currently most companies are under no obligation to disclose their

Interventions

algorithms or internal processes, therefore any changes to these processes will not be obvious to outsiders. Second it is hard to quantify ethics and bias, these terms describe social phenomena which are difficult to observe. However, the association will have to determine major societal biases, such as racial, gender, ableist, and ageist bias, and have a way to measure new AI technology against determined metrics. These metrics will vary depending on the context and technology, which further emphasizes the need for the association to be led by industry experts.

It is important when measuring the impact of the association to not only focus on the outcomes of the AI technologies, but also their processes. Bias in technology is largely built in through internal practices such as data collection (Kantayya, 2020), and lack of diversity among tech employees.

Having metrics which measure internal processes, such as diversity among the data pools used for machine learning could have a measurable impact on the outcome of the technology. Rather than regulate these internal processes directly, the association could provide ethical frameworks or 'best practices' that would support companies in achieving these goals.

Limitations

A common problem with internal ethics review panels is that they do not have real power to shutdown projects or implement policies and they are, therefore, largely ineffectual (Johnson & Lichfield, 2019; Hao, 2021). Individual companies will inevitably prioritize corporate interests over the ethics panel recommendations if they are at odds.

There would have to be real consequences for failing to correct, change, or remove AI technologies, which were deemed egregious by the association. This can be countered in part by ensuring influential figures and industry leaders participate in the creation of this association to lend their expertise and authority to the project. This could also be a place for governments or intergovernmental groups to be stakeholders in the formation of this type of association and impose policies, which would have tangible consequences for unethical practices.

A second limitation is the industry culture and algorithmic intellectual property. Technology companies are very secretive when it comes to algorithms and the decision-making processes behind their AI (Pasquale, 2017; Cofone & Strandburg, 2019; Kilburn, 2021). Society has access to the finished product, but the algorithm itself is protected intellectual property. The how’s and why’s of a decision made by a piece of AI technology are not available to anyone wishing to study or evaluate them. This level of secrecy would have to change to be able to implement this intervention.

Protecting the IP of the companies in question is still important. There would have to be ways to evaluate the AI product for biases and still protect the IP. Policies such as the ones used to evaluate agribusiness, or pharmacy products are good examples of industries with similar concerns (Government of Canada, 2021).

Interventions

Though this intervention seems challenging, there are numerous examples of associations and organizations in different spaces, which operate in a similar manner to protect consumers and ensure ethical practices, including journalism, advertising, medicine, biology, genetics, robotics, and mining. These types of associations are imperative to ensuring future ethical research and development in these spaces, and a subject as vast and complex as AI technology will need a similar association for the future.

There is already a push for this type of oversight from members within and outside of the industry. (Jordan, 2019; Ghaffary & Kantrowitz, 2021).

Interventions

INTERVENTION 4: PUBLIC EDUCATION AND AWARENESS

Level: Macro/Societal

Stage: Not applicable. A parallel intervention.

Key Actors: General public, media.

Timeline: 1-3 years. A campaign can be put together quickly but as a passive intervention, repetition of the campaign will be needed to maximize impact.

Leverage Points: Media, STEM and tech literacy initiatives, advocacy groups and industry watchdogs

Support Processes: Governmental expert advisory councils, similar public education and awareness campaigns

Public Awareness Campaign
Layer: Macro/Societal

Intervention: Public awareness campaign on emerging technology biases

Context

Societal technology biases and lack of awareness of understanding of how emerging technologies work or their implications are key barriers to changing the technology ecosystem to mitigate bias. Currently, two pervasive technology biases persist in society: the belief in *technobenevolence*, which is a societal belief in the purity of technology, and the *new technology bias*, a bias which causes people to favour emerging technologies due to their novelty and particularly if endorsed by experts (Elsbach & Stigliani, 2020; University of Toronto Joint Centre for Bioethics, 2019). These biases create reinforcing loops that that hinder meaningful interrogation of AI and its algorithms.

An example of this is the case of now-defunct health technology company, Theranos, who claimed it could conduct cheaper, faster, and more efficient blood tests. As a former Silicon Valley start-up, it was once valued at over $9 billion and through its high-profile board members and elite affiliations, the company was able to raise $700 million despite a lack of transparency around its product or the business. This persisted, endangering the health of users, until the problems were reported in the news (Elsbach & Stigliani, 2020; Waltz, 2017).

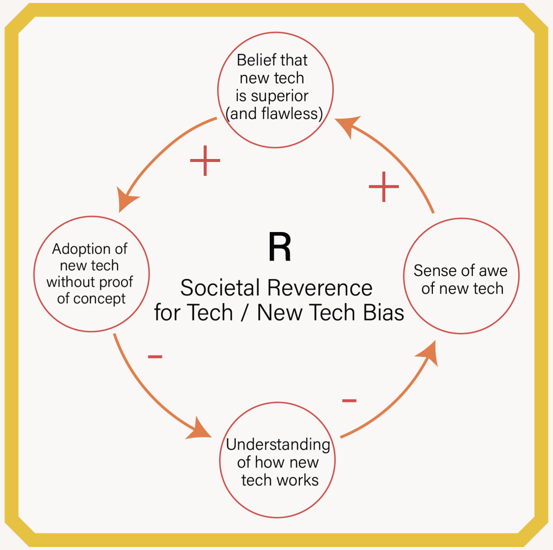
This is an example of the new technology bias where individuals are more likely to believe in the value and effectiveness of a technology if endorsed by a perceived expert (Elsbach & Stigliani, 2020).

It perpetuates a belief that emerging technologies are the domain of experts and allows non-experts to side-step the responsibility of investigating the echnologies so that it ends up being adopted without due public scrutiny and validation.

Interventions

It can also mask the fact that the experts developing the AI are just a subset of the general of the population separated by "the exclusive culture of computing education and industry" (Bogost, 2019), and equally subject to their own unconscious biases.

As a consequence of these societal beliefs, there is little incentive for technology companies to alter their approach. Even more so, if altering their approach could introduce a delay in their product lifecycle that would hinder their speed to market. As quantitative futurist, author and professor at New York University, Amy Webb, says "a moral imperative is unlikely to motivate public companies" (Bogost, 2019). Public pressure from a more informed society has the potential to create a strong incentive for companies to change if they feel their profits are endangered.


Belief that new tech is superior (and flawless)
+
+
Adoption of new tech without proof of concept
R
Societal Reverence for Tech / New Tech Bias
Sense of awe of new tech
–
–
Understanding of how new tech works

Strategy Outline

In parallel to the early, mid, and late-stage interventions, a complementary large-scale public awareness and education intervention is recommended to challenge common myths and misconceptions of AI and identify risks to be investigated. Its primary leverage is using the power of knowledge in the hands of the public to create pressure on government and technology companies to develop policies and practices for ethical AI. Evidence from other education campaigns demonstrates that increasing the flow of information to the public is an effective way to increase awareness and discussion of the topic and prompt policy change, particularly when implemented in tandem with other interventions (Centers for Disease Control & Prevention, 2020; Wakefield et al, 2010).

As the general public has a low ability to directly impact bias in AI technology, a large-scale campaign with the broadest reach to increase general discourse would maximize impact. The campaign would use digital marketing, PR, and partnerships and networking to build support and increase the audience reach. A campaign would also help support the introduction of balancing loops and delays into the system. It has the capacity to reach other actors as part of the general public and influence their thoughts and actions in other levels of the system. It also would plant the seeds for a broader, longer lasting paradigm shift by bringing attention to society's current technological biases.

Interventions

Implementation

The campaign would take a multi-pronged approach using digital marketing efforts, traditional PR tactics, and community partnership and networking to reach the widest audience. Messaging would centre around providing insights into:

1. **The Problem:** Cultivate awareness of how bias is being embedded in AI
2. **The Implications:** Educate on the consequences of implicit biases
3. **The Breadth:** Highlight the ubiquity of AI who this impacts—that is, the majority of the general public, who interact with AI on daily basis
4. **Potential Solutions:** Present pathways to creating change through other interventions.

Annual Campaign

An annual campaign would be run over a specific timeframe each year over multiple years, ideally in alignment with a national or global awareness day. It would use a digital presence—website, social media—to launch the campaign and direct action.

Use of strong visual identity and impactful imagery will strengthen impact and retention of messages. Additional tactics to increase engagement could include an interactive audience challenge—such as the ALS Ice Bucket challenge, Movember, or the No Make-Up Selfie—and public workshops to gauge the perceptions and concerns around AI.

The Government of Canada's Advisory Council on Artificial Intelligence, which has a Public Awareness Working Group, is already reaching out to Canadians with the intent to "foster trust in AI" and could be a beneficial partner in developing the campaign (Government of Canada, 2021).

Partnership & Collaboration

Partners and collaborations also have the potential to increase the reach and legitimacy of the campaign. With their existing networks, reach and established reputations, their support would help promote the issue and help build its credibility.

There are a variety of existing initiatives in the STEM, tech diversity, and tech literacy spaces, like the Advisory Council on AI or organizations like Canada Learning Code, who could be identified and approached. Another opportunity would be to approach universities and colleges with technology programs and technology bootcamps to offer workshops and resources to encourage ethical thinking of AI early.

Measuring Impact

Measurement is important to identify the relative success of any initiative, however, this can be challenging for public education and awareness campaigns (Wakefield et al., 2010). In this intervention, measurement around participation, through digital metrics—such as webpage visits, social media engagement, use of campaign hashtag, sentiment tracking, etc.—and increased community collaboration through partnerships and recruitment of volunteers will be important metrics in quantifying the impact of the campaign. In the longer term, increased demand for ethical and regulated AI technology would also be an indicator of successful impact.

Limitations

Education campaigns are passive in their approach to change as they rely on their audiences to act on the messages they present (Wakefield et al., 2010).

Interventions

in this instance, this would be no different, requiring that this intervention be able to convince the public that this issue is relevant, urgent, and has harmful consequences. Case studies, compelling numbers, and heartfelt impact stories which make visible the consequences of biased algorithms are important elements to creating a compelling campaign and motivating members of the public to action.

Similarly, education campaigns can be limited in impact on their own, particularly in this case, as the general public has no direct access to changing the system (Wakefield et al., 2010). This is in part because of the cluttered media environment the content is published into (Nielsen et al., 2016). Therefore, the intent of this approach is not to launch it on its own, but as part of a series of interventions that address different aspects of the system. An essential part of developing and launching an impactful campaign would be to set clear campaign goals and calls-to-action, conduct thorough audience research to best target the messaging, develop consistent imagery for quick recognition, and explore partnership opportunities to leverage existing allies and networks.

Discussion

In our ideal future, technology is used responsibly, for good. The process of its creation is democratic and transparent and protects the public, particularly vulnerable groups. To achieve this, more voices and regulations are needed, as bias is pervasive throughout the development lifecycle.

Our interventions show many elements of the pathway to change, such as public awareness, government regulations, ethics boards, incentives for social good, and diversity. These are most impactful if implemented together. Each one strengthens the others and builds momentum that will change how the AI system operates. They also tap into select actor's power, such as the government and media, balancing the power asymmetry that currently favours AI developers.

These interventions need to be sustained to ensure the system does not revert to its present state. They are integral to introducing balancing loops and delays into the system to balance the current actions where biases and dominant powers reinforce each other. They engage different levels of the system—from parameters all the way to a paradigm shift—which creates a robust approach to affecting change (Meadows, 1999). Given the relative speed at which the different intervention levels are able to implement change, it also allows for short, mid and long-term options to continue the momentum.

Additional research could be conducted to further deepen the understanding of nuances within the system and within specific industries such as healthcare and the criminal justice system. Each intervention operates within its own ecosystem of stakeholders and so further research would reveal additional insight and differences within these subsystems which include the financial/investment, government, business development, and labour systems.

Researching adjacent socio-technical systems—such as STEM education, the media, innovation hubs—can also enhance the nuance and understanding of the broader system. Further research on the development lifecycles and within different sized technology companies to identify unique needs and insights that can help make these interventions more robust.

Appendix A: Case Study - COMPASS

OVERVIEW

COMPAS (short for Correctional Offender Management Profiling for Alternative Sanctions) owned by Equivant technologies. It is a 'risk assessment' AI technology which uses an algorithm to predict the likeliness of recidivism by a defendant (Larson et al., 2016).

COMPAS assesses Pretrial Release Risk, General Recidivism and Violent Recidivism using an undisclosed algorithm (Northpointe, 2015) and provides a defendant with a 'score' for each of these categories. Judges then use this score to help inform their sentencing decisions (Angwin et al., 2016).

The technology that the COMPAS risk assessment tool uses to make its decisions is based on an undisclosed algorithm. In order to make its assessment, the technology needs information about a person, which it inputs into the system. There are several information points requested about the individual such as age, the neighborhood they live in, employment status, previous offences, and whether they have family members who have been incarcerated. It is worth noting that none of the information put into the algorithm is specifically about race.

Based on a large ProPublica study of 7000 individuals, COMPAS reliably predicts the likelihood of an individual reoffending roughly 60% of the time (Angwin et al., 2016). **However, the technology is more likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants** (Angwin et al., 2016). Based on their statistical analysis this disparity was not based on prior crimes or the type of crime they were arrested for (Angwin et al., 2016). This racial disparity can be explored by looking at the system through the framework developed in this report.

Using the Intervention Strategies to "unbias" this Tech

Funding: COMPAS is owned by Equivant Technologies, which is a large publicly traded corporation. Their funding structure relies on shareholders and raising capital through traditional banking avenues to fund their ventures. This type of funding is not in and of itself bias, however it does instigate a push for technologies with high earning potential, such as ones targeted to the American criminal justice system.

If the company raised capital with ethical investors, the potential harmful outcomes of their technology would be an important factor before receiving financing for new AI technologies.

Hiring & Diversity: It is not possible to know the composition of the team which designed and developed COMPAS specifically. If we look at the industry average the number of black employees is anywhere from 3%-10% (Dickey, 2019). Including more diverse employees in the development of this technology could have meant that the racial disparity evident in the results would have been discovered sooner.

While none of the inputs that the algorithm uses to determine its score are specifically about race, many of them are a proxy for race when taking into account the history of race and racial bias in American culture. For example, there has historically been a bias in policing of black neighborhoods, these neighborhoods can be overpoliced with the inhabitants being cited or charged, where a white counterpart would not be (Weir, 2016).

Appendix A: Case Study - COMPASS

This means that black individuals might be more likely to have family members with previous recorded offences or live in a neighborhood with more individuals with a criminal record. This is something that having a diverse team can help bring to light, by having more experience with racial disparities in America.

Ethics Review Association: If COMPAS had to apply to an ethics association and receive an evaluation for this technology the blatant racial disparity in the output of the technology would most likely have been flagged. Furthermore, the question about the value and deployment of an AI technology of this kind might have been called into question.

Public Awareness Campaign: Finally, if there was recognition from the public about the use of AI technologies such as this one, a number of things might be different. The blind faith that the judges, sentencing, and parole officers have in these types of technologies would likely be different. There might also be a push back from the public in the use of these technologies, stronger public awareness might also cause individual defendants to question the use of this type of tech in their sentencing.

Appendix B: Case Study – Apple Card

OVERVIEW

In 2019, Apple and Goldman Sachs teamed up to launch a credit card called *Apple Card* which could be used across all Apple devices.

Shortly after it was released multiple reports from couples about the gender disparity of the Apple Card were brought to light. After applying for Apple Card with the same assets and similar or in some cases higher credit scores than their husbands, women would be approved for credit limits 2x - 30x lower than their spouses (Vigdor, 2019).

Even Apple's cofounder Steve Wozniak highlighted that he received 10 times the limit his wife had when they applied for the Apple Card (Wieczner & Morris, 2019).

Goldman Sachs emphasized that gender was not one of the inputs for the algorithm, and that it uses many other inputs such as credit score, and assets, to assess "creditworthiness." Furthermore, the bank explained that they had used a third-party company to audit the technology (Knight, 2019).

This gender disparity can be explored by looking at the system through the framework developed in this report.

Using the Intervention Strategies to "unbias" this Tech

Funding: Apple Card is funded by Apple, and financially backed by the bank Goldman Sachs. The motivating factor of a piece of technology such as this one would be to increase the bank revenue (for Goldman) and increase the ubiquity of Apple products in users lives (for Apple).

Hiring and Diversity: At Apple women make up 33% of the total workforce, that number drops to 29% in leadership roles and 23% in tech roles (Richter, 2020). Similar to proxies for race, many inequities exist in banking that historically disenfranchised women. There are also many changes that happen during marriage which disproportionately affect women and potentially their finances. Such as changing their last name or closing their bank account to join a spouse's financial institution.

These are considerations which women on a design or development team would be more likely to highlight as potential places where bias might creep into an algorithm inadvertently.

Ethics Review Association: If Apple Card had had to apply to an ethics review panel for scrutiny before the release of this product there is a chance that some of the discrepancy in credit approvals based on gender would have been flagged before the product was launched. As it stands the Apple Card was reviewed by the New York State Department of Financial Services (Vigdor, 2019). It is worth noting however that the very fact that the algorithm does not use gender as an input means that reviewing the data happens without gender discrimination being taken into account (Knight, 2019).

Public Awareness: The knowledge that these typed of technologies can be and often are biased based on their algorithms is important for the public to recognize. The push back that Apple and Goldman Sachs received after the launch of this technology is notable, partially because the bias was almost immediately noticed and flagged as problematic. This seems to rarely happen when it comes to AI technology.

References

Ad Standards. (n.d). About us. https://adstandards.ca/about/

American Psychological Association. (2008). Public education campaign overview. APA.org. https://www.apa.org/practice/programs/campaign

Appen. (2020, July 2). How to reduce bias in AI. Top eight ways to overcome and prevent AI bias. Appen.com. https://appen.com/blog/how-to-reduce-bias-in-ai/#:~:text=To%20minimize%20bias%2C%20monitor%20for,initiative%20will%20ultimately%20fall%20apart

Angwin, J., Larson, J., Mattu, S., & Kirchner, L., (2016, May 23). *Machine bias there's software used across the country to predict future criminals. And it's biased against blacks.* Pro Publica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Anyoha, R. (2017, August 28). The history of artificial intelligence. *Science in the News.* https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Artificial Intelligence. (n.d). Builtin. Retrieved on April 16, 2020, from https://builtin.com/artificial-intelligence

Atkinson, R. D. (2014). Understanding the U.S. national innovation system (SSRN Scholarly Paper ID 3079822). *Social Science Research Network*. https://doi.org/10.2139/ssrn.3079822

Babich, N. (2019, October 11). Everything you need to know about beta testing. Xd Ideas. Adobe. https://xd.adobe.com/ideas/process/user-testing/everything-you-need-to-know-about-beta-testing/

Bias. (n.d.). Psychology Today. Retrieved February 18, 2021, from https://www.psychologytoday.com/ca/basics/bias

Blackman, R. (2020, October 15). A practical guide to building ethical AI. *The Harvard Business Review*. https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai

Bogost, I. (2019, June 25) The problem with diversity in computing. *The Atlantic*. https://www.theatlantic.com/technology/archive/2019/06/tech-computers-are-bigger-problem-diversity/592456/

Braun, W. (2002). The System Archetypes.

Bueschen, R. (2015, September 24). The surprising bias of venture capital decision-making. *TechCrunch.* https://social.techcrunch.com/2015/09/24/the-surprising-bias-of-venture-capital-decision-making/

Buolamwini, J. (2019, March 29). *The coded gaze: Bias in artificial intelligence | Equality Summit.* [Video]. YouTube. https://www.youtube.com/watch?v=eRUEVYndh9c

Cambridge Dictionary. (n.d). Algorithm. In Cambridgedictionary.com. Retrieved on April 16, 2021, from https://dictionary.cambridge.org/dictionary/english/algorithm

Cambridge Dictionary. (n.d). Users. In Cambridgedictionary.com. Retrieved on April 16, 2021, from https://dictionary.cambridge.org/dictionary/english/user

Chae, T. (2019, May 1). *How do VCs evaluate early stage startups versus later stage ones?* Forbes. https://www.forbes.com/sites/quora/2019/05/01/how-do-vcs-evaluate-early-stage-startups-versus-later-stage-ones/?sh=806c86c3b3e2

Centers for Disease Control and Prevention. (2020, October 15). About the campaign. Centers for Disease Control and Prevention. https://www.cdc.gov/tobacco/campaign/tips/about/index.html

Chakravorti, B. (2020, December 4). To increase diversity, U.S. tech companies need to follow the talent. *Harvard Business Review*. https://hbr.org/2020/12/to-increase-diversity-u-s-tech-companies-need-to-follow-the-talent

Cofone, I., Strandburg, K. (2019, June). Strategic games and algorithmic secrecy. *McGill LJ, (64)* 4, 623 https://lawjournal.mcgill.ca/article/strategic-games-and-algorithmic-secrecy/

Copeland, B. J. (n.d). Artificial intelligence. In *Britannica.com dictionary*. Britannica. Retrieved on April 16, 2021, from https://www.britannica.com/technology/artificial-intelligence

Crowther-Heyck, H. (2008). Defining the computer: Herbert Simon and the bureaucratic mind—part 1. *IEEE Annals of the History of Computing*, *30*(2), 42–51. https://doi.org/10.1109/MAHC.2008.18

Curry, A., & Hodgson, A. (2008). Seeing in multiple horizons: Connecting futures to strategy. *Journal of Futures Studies*, *13*(1), 20.

Data science and machine learning. (2020, August 27). IBM. https://www.ibm.com/analytics/machine-learning

DeNisco Rayome, A. (2018, February 7). *5 eye-opening statistics about minorities in tech*. TechRepublic. https://www.techrepublic.com/article/5-eye-opening-statistics-about-minorities-in-tech/

Dick, S. (2019). Artificial intelligence. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.92fe150c

Dickey, M. (2019, June 17). *The future of diversity and inclusion in tech*. TechCrunch. https://techcrunch.com/2019/06/17/the-future-of-diversity-and-inclusion-in-tech/

References

Downy, L. (2019, October 29). Algorithm. *Investopedia*. Retrieved on April, 16, 2021, from https://www.investopedia.com/terms/a/algorithm.asp

Elsbach, K.D., & Stigliani, I. (2020, September 23). Evaluating new technology? You're more biased than you may realize. *MIT Sloan Management Review.* https://sloanreview.mit.edu/article/evaluating-new-technology-youre-more-biased-than-you-may-realize/

Employee resource groups. (2021). Catalyst. https://www.catalyst.org/topics/ergs/

European Commission. (2020). *Artificial intelligence a European approach to excellence and trust* [White paper]. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Fassin, Y., & Drover, W. (2017). Ethics in entrepreneurial finance: Exploring problems in venture partner entry and exit. *Journal of Business Ethics*, *140*(4), 649–672.

Fernández Campbell, A. F. (2019, April 23). *Google employees say the company is punishing them for their activism*. Vox. https://www.vox.com/2019/4/23/18512542/google-employee-walkout-organizers-claim-retaliation

Ferrary, M., & Granovetter, M. (2009). The role of venture capital firms in Silicon Valley's complex innovation network. *Economy and Society*, *38*(2), 326–359. https://doi.org/10.1080/03085140902786827

Frankenfield, J., Scott, G. (2021, March 8). Artificial intelligence. *Investopedia*. Retrieved on April 16, 2021, from https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp

Fundr. (n.d.). Fundr. Retrieved March 29, 2021, from https://www.fundr.ai

Gassam Asare, J. (2018, September 26). *How to retain diverse talent*. Forbes. https://www.forbes.com/sites/janicegassam/2018/09/26/how-to-retain-diverse-talent/?sh=7ded8e5e2d33

Ghaffary, S., Kantrowitz, A. (Hosts). (2021, March 16). A military contract tests Google's open culture (3(5). [Audio Podcast Episode]. In *Land of The Giants*. The Google Empire. Vox Media. https://megaphone.link/VMP9674926755

Gharajedaghi, J. (2004). *Systems methodology: A holistic language of interaction and design seeing through chaos and understanding complexities*.

Gibney, E. (2020, January 24). The battle for ethical AI at the world's biggest machine-learning conference. Bias and the prospect of societal harm increasingly plague artificial-intelligence research — but it's not clear who should be on the lookout for these problems. *Nature*. https://www.nature.com/articles/d41586-020-00160-y

Global Impact Investing Network. (2011). *Impact-based incentive structures*. https://thegiin.org/assets/documents/pub/impact-based-incentive-structures-aligning-fund-manager-comp.pdf

Global Impact Investing Network. (2018, March). *Roadmap for the future of impact investing: Reshaping financial markets*. https://thegiin.org/assets/GIIN_Roadmap%20for%20the%20Future%20of%20Impact%20Investing.pdf

Green, D. [Centre for Ethics]. (2019, Nov 27). Ethics of AI in context: Making ethics in machine learning. [Video]. YouTube. https://www.youtube.com/watch?v=AiobQhZSE94

Gonfalonieri, A. (2020, July 20). *What is an AI algorithm?* Medium. https://medium.com/predict/what-is-an-ai-algorithm-aceeab80e7e3

Goodman, M., Kemeny, J., & Roberts, C. (2000). *The language of systems thinking: "Links" and "loops."*

Government of Canada. (2021, March 8). Advisory Council on Artificial Intelligence. Retrieved April 16, 2021, from http://www.ic.gc.ca/eic/site/132.nsf/eng/home

Government of Canada. (2021, March 29). Canada's regulatory system for foods with health benefits - An overview for industry. Retrieved April 16, 2021, from https://www.agr.gc.ca/eng/canadas-agriculture-sectors/food-products/processed-food-and-beverages/trends-and-market-opportunities-for-the-food-processing-sector/canada-s-regulatory-system-for-foods-with-health-benefits-an-overview-for-industry/?id=1274467299466

Hao, K. (2018, November 17). *What is machine learning?* MIT Technology Review. https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/

Hao, K. (2019, April 15). *Congress wants to protect you from biased algorithms, deepfakes, and other bad AI.* MIT Technology Review. https://www.technologyreview.com/2019/04/15/1136/congress-wants-to-protect-you-from-biased-algorithms-deepfakes-and-other-bad-ai/

Hao, K. (2021, March 11). *How Facebook got addicted to spreading misinformation.* MIT Technology Review. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/

Harrison, S. (2019, October 1). Five years of tech diversity reports—And little progress. *Wired*. https://www.wired.com/story/five-years-tech-diversity-reports-little-progress/

Harroch, R., & Sullivan, M. (2018, March 29). *A guide to venture capital financings for startups*. Forbes. https://www.forbes.com/sites/allbusiness/2018/03/29/a-guide-to-venture-capital-financings-for-startups/

References

Hauser, R. [TED Institute]. *Can we protect AI from our biases?* [Video]. YouTube. https://www.youtube.com/watch?v=eV_tx4ngVT0

Hawkins, A. J. (2018, May 9). *Inside the lab where Waymo is building the brains for its driverless cars.* The Verge. https://www.theverge.com/2018/5/9/17307156/google-waymo-driverless-cars-deep-learning-neural-net-interview

Henton, D., & Held, K. (2013). The dynamics of Silicon Valley: Creative destruction and the evolution of the innovation habitat. *Social Science Information*, *52*(4), 539–557. https://doi.org/10.1177/0539018413497542

Heyck, H. (2005). *Herbert Simon: The bounds of reason in modern America*. Johns Hopkins University Press.

Huffman, D., & Quigley, J. M. (2002). The role of the university in attracting high tech entrepreneurship: A Silicon Valley tale. *The Annals of Regional Science*, *36*(3), 403–419. https://doi.org/10.1007/s001680200104

Hutson, M. (2021, Febuary, 15). Who should stop unethical A.I.? At artificial-intelligence conferences, researchers are increasingly alarmed by what they see. *The New Yorker. https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai*

IBM Cloud Education. (2020, June 3). Artificial intelligence. *IBM*. Retrieved on April 16, 2021, from https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

Introducing the Systemic Design Toolkit at RSD7. (2018, November 12). Systemic Design Toolkit. https://www.systemicdesigntoolkit.org/news/introducing-the-systemic-design-toolkit-at-rsd7

Johnson, B., Lichfield, G. (2019, April 6). *Hey Google, sorry you lost your ethics council, so we made one for you.* MIT Technology Review. https://www.technologyreview.com/2019/04/06/65905/google-cancels-ateac-ai-ethics-council-what-next/

Jordan, S. (2019, October). *Designing an artificial intelligence research review committee.* Future of Privacy Forum. https://fpf.org/wp-content/uploads/2019/10/DesigningAIResearchReviewCommittee.pdf

Koenig, A. (2016, July 19). *Governments in social impact investing – What's next on the global agenda?* IBAN. https://www.inclusivebusiness.net/ib-voices/governments-social-impact-investing-whats-next-global-agenda

Kantayya, S. (Producer). Kantayya, S. (Director).
(2020). *Coded Bias* [Film]. https://www.netflix.com.

Kaushal, A., Altman, R., & Langlotz, C. (2020, November 17). *Health Care AI Systems Are Biased*. Scientific American. https://www.scientificamerican.com/article/health-care-ai-systems-are-biased/

Kenyon, M. (2020, September 1). *Algorithmic Policing in Canada Explained*. The Citizen Lab. https://citizenlab.ca/2020/09/algorithmic-policing-in-canada-explained/

Kilburn, T. [RSA] (2021, February 18). *Tech companies shroud their algorithms in secrecy. It's time to pry open the black box*. [Video]. Aeon. https://aeon.co/videos/tech-companies-shroud-their-algorithms-in-secrecy-its-time-to-pry-open-the-black-box

Kirkwood, C.W. (1998). System behavior and causal loop diagrams. In *System Dynamics Methods: A Quick Introduction.* CC BY-NC 3.0 http://www.public.asu.edu/~kirkwood/sysdyn/SDIntro/ch-1.pdf

Knight, W. (2019, November 19). The Apple Card didn't 'see' gender—and that's the problem. *Wired.*

Lau, J. (2020, September 3). *Google Maps 101: How AI helps predict traffic and determine routes*. Google. https://blog.google/products/maps/google-maps-101-how-ai-helps-predict-traffic-and-determine-routes/

Li, M. (2020, October 26). To build less-biased AI, hire a more diverse team. *Harvard Business Review.* https://hbr.org/2020/10/to-build-less-biased-ai-hire-a-more-diverse-team

Machine Learning: What it is and why it matters. (2021). SAS. https://www.sas.com/en_us/insights/analytics/machine-learning.html

Manyika, J., Presten, B., & Silberg, J. (2019, October 25). What do we do about the biases in AI? *Harvard Business Review.* https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Marr, B. (2019, December 16). *The 10 best examples of how AI is already used in our everyday life*. Forbes. https://www.forbes.com/sites/bernardmarr/2019/12/16/the-10-best-examples-of-how-ai-is-already-used-in-our-everyday-life/

Martin, N. (2019, September 30). *Artificial Intelligence Is Being Used To Diagnose Disease And Design New Drugs*. Forbes. https://www.forbes.com/sites/nicolemartin1/2019/09/30/artificial-intelligence-is-being-used-to-diagnose-disease-and-design-new-drugs/

McCarthy, J. (2007, November 12). What is artificial intelligence? *Stanford University*. http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

Meadows, D. (1999). Leverage points: Places to intervene in a system. The Sustainability Institute.

References

Mentoring: Benefits, challenges, and new approaches. (2011, June). Diversity Best Practices. https://www.diversitybestpractices.com/sites/diversitybestpractices.com/files/import/embedded/anchors/files/_attachments_articles/rr-mentoringjune2011.pdf

Merriam-Webster. (n.d.). Algorithm. In *Merriam-Webster.com dictionary*. Retrieved April 16, 2021, from https://www.merriam-webster.com/dictionary/algorithm

Mozilla Foundation. (2020, December). *Creating trustworthy AI*. [White paper]. https://assets.mofoprod.net/network/documents/Mozilla-Trustworthy_AI.pdf

Nielsen, R. K., Cornia, A., & Kalogeropoulos, A. (2016). Challenges and opportunities for news media and journalism in an increasingly digital, mobile, and social media environment. *Council of Europe Report DGI*, *2016*(18), 41. https://rm.coe.int/16806c0385

Newell, A., & Simon, H. (1972). *Human problem solving*. Prentice-Hall.

Northpointe, I. (2015). Practitioner's guide to COMPAS Core. https://assets.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf

NVCA, Venture Forward, & Deloitte. (2021). *VC human capital survey—Third edition* (p. 32). https://nvca.org/research/nvca-deloitte-human-capital-survey/

OECD. (n.d). Artificial intelligence. Retrieved April 16, 2021, from https://www.oecd.org/going-digital/ai/#:~:text=The%20OECD.AI%20Policy%20Observatory,over%2060%20countries%20and%20territories

O'Neil, C. (2017, April). *Cathy O'Neil: The era of blind faith in big data must end.* [Video]. https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end

O'Neil, C. [RSA]. (2018, April). *The Truth About Algorithms.* [Video]. YouTube. https://www.youtube.com/watch?v=heQzqX35c9A

Pasquale, F. (2017, June 1). *Secret algorithms threaten the rule of law.* MIT Technology Review. https://www.technologyreview.com/2017/06/01/151447/secret-algorithms-threaten-the-rule-of-law/

PCV InSight & The Initiative for Responsible Investment. (2014). *Impact investing policy in 2014: A snapshot of global activity*. https://iri.hks.harvard.edu/files/iri/files/impact_investing_policy_in_2014_a_snapshot_of_global_activity_2014_full_report.pdf

Richter, F. (2020, February 19) GAFAM: Women still underrepresented in tech. *Statista*. https://www.statista.com/chart/4467/female-employees-at-tech-companies/

Rosenbaum, E. (2019, September 26). *Liberal arts degree? No degree at all? You are the perfect candidate for a tech job*. CNBC. https://www.cnbc.com/2019/09/26/tech-jobs-now-a-fit-for-a-liberal-arts-degree-or-no-degree-at-all.html

Schwartz, A., & Finighan, R. (2020, July 17). Impact investing won't save capitalism. *Harvard Business Review*. https://hbr.org/2020/07/impact-investing-wont-save-capitalism

Samuel, S. (2019, March 6). *A new study finds a potential risk with self-driving cars: Failure to detect dark-skinned pedestrians.* Vox. https://www.vox.com/future-perfect/2019/3/5/18251924/self-driving-car-racial-bias-study-autonomous-vehicle-dark-skin

Senge, P.M. (2006). *The fifth discipline.* Crown Business.

Silberg, J., & Manyika, J. (2019). *Notes from the AI frontier: Tackling bias in AI (and in humans)*. McKinsey Global Institute. 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.pdf

Smith, C. (2019, November 19). Dealing with bias in artificial intelligence. *The New York Times.* https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html

Stanborough, R. J. (2020, May 28). *Cognitive bias: Understanding how it affects your decisions*. Healthline. https://www.healthline.com/health/mental-health/cognitive-bias

Stanford University. (n.d.). Data USA. Retrieved April 8, 2021, from https://datausa.io/profile/university/stanford-university

Systemic Design Toolkit. (2021). Retrieved April 16, 2021, from https://www.systemicdesigntoolkit.org/

Taleb, N. N., & Sandis, C. (2014). The skin in the game heuristic for protection against tail events. *Review of Behavioral Economics*, *1*, 1–21.

Tepper, N. (n.d.). *Meet Fundr, the startup that wants to use AI to take on angel investors' biases*. Built In Austin. Retrieved March 26, 2021, from https://www.builtinaustin.com/2020/11/24/fundr-launch-diversify-angel-investment-founders

Tiell, S. (2019, November 15). Create an ethics committee to keep your AI initiative in check. *Harvard Business Review.* https://hbr.org/2019/11/create-an-ethics-committee-to-keep-your-ai-initiative-in-check

References

University of California-Berkeley. (n.d.). Retrieved April 8, 2021, from https://datausa.io/profile/university/university-of-california-berkeley

University of Toronto Joint Centre for Bioethics. (2019, November 29). *Assessing risk, automating racism: Reimagining the default settings of technology in healthcare.* [Video]. YouTube. https://www.youtube.com/watch?v=QXr9AovMoKk

Uzzi, B. (2020, November 4). A simple tactic that could help reduce bias in AI. *Harvard Business Review.* https://hbr.org/2020/11/a-simple-tactic-that-could-help-reduce-bias-in-ai

Vigdor, N. (2019, November 10). Apple Card investigated after gender discrimination complaints. *The New York Times.* https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

Vincent, J. (2019, April 3). The problem with AI ethics. Is Big Tech's embrace of AI ethics boards actually helping anyone? *The Verge.* https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech

Vinney, C. (2018, October 21). *How cognitive biases increase efficiency (and lead to errors).* ThoughtCo. https://www.thoughtco.com/cognitive-bias-definition-examples-4177684

Wakefield, M. A., Loken, B., & Hornik, R. C. (2010). Use of mass media campaigns to change health behaviour. *Lancet, 376*(9748), 1261–1271. https://doi.org/10.1016/S0140-6736(10)60809-4

Walch, K. (2019, July 26). *The Growth Of AI Adoption In Law Enforcement*. Forbes. https://www.forbes.com/sites/cognitiveworld/2019/07/26/the-growth-of-ai-adoption-in-law-enforcement/

Walch, K. (2020, April 5). *Why AI Is Transforming The Banking Industry*. Forbes. https://www.forbes.com/sites/cognitiveworld/2020/04/05/why-ai-is-transforming-the-banking-industry/

Waltz, E. (2017). After Theranos. *Nature Biotechnology, 35*(1), 11–16. https://doi.org/10.1038/nbt.3761

Weinberger, M., Stuart, M., & Protin, C. (2020, December 18). *Animated timeline shows how Silicon Valley became a $2.8 trillion neighborhood*. Business Insider. https://www.businessinsider.com/silicon-valley-history-technology-industry-animated-timeline-video-2017-5

Wieczner, J., Morris, D. (2019, November 13). The Apple Card's algorithm goes both ways on women's credit limits. *Fortune*. https://fortune.com/2019/11/13/apple-card-bias-women-goldman-sachs/#:~:text=By%20Saturday%20afternoon%2C%20the%20New,bank%E2%80%94had%20a%20gender%20bias.

West, S. M., Whittaker, M., & Crawford, K. (2019). *Discriminating systems: Gender, race, and power in AI (p. 33).* AI Now Institute. https://ainowinstitute.org/discriminatingsystems.pdf

Weir, K. (2016, December). Policing in black and shite. *American Psychological Association 47*(11), 33. https://www.apa.org/monitor/2016/12/cover-policing

Wilson, B., Hoffman, J., Morgenstern, J. (2019, February 17). Predictive inequity in object detection. https://arxiv.org/pdf/1902.11097.pdf

Woo, M., McIntosh, K., & Stanley-McAulay, D. (2018, May 7). *How to plug the leaky bucket: Retention strategies for maintaining a diverse workforce*. EDUCAUSE. https://er.educause.edu/articles/2018/5/how-to-plug-the-leaky-bucket-retention-strategies-for-maintaining-a-diverse-workforce

Zhou, M. (2019, August 22). *ESG, SRI, and impact investing: What's the difference?* Investopedia. https://www.investopedia.com/financial-advisor/esg-sri-impact-investing-explaining-difference-clients/