Power and Progress in Algorithmic Bias

—a landscape analysis—

Written by:
Kristine Gloria, Ph.D.

Table of Contents

Executive Summary

The following presents the current, evolving landscape of the study of algorithmic bias. The purpose of this work is to highlight key learnings over the past decade, reviewing how we got here and considering where we can and should go from here. The work is informed by a multidisciplinary study of algorithmic bias literature, illuminating the many critiques, harms, challenges, and opportunities in this line of inquiry. This text is not intended to be a comprehensive, exhaustive literature review. The research and work in this space is abundant and any attempt to capture all of it would succeed only in underrepresenting its vastness and value. Instead, readers of this work should consider each report section as a provocation for action.

The overall work is motivated by a need to understand how technology, and information culture, can help eradicate social inequalities.[1] Particularly, we need to understand the harmful impact of algorithmic bias on marginalized communities. These groups have not only been systematically omitted from the data and the development of automated decision-making systems throughout history but have also been actively harmed by such systems. As with other socio-technical critiques, evidence of algorithmic bias in systems underscores the theory that human values and human decisions are embedded in the technology. Author and researcher Brian Christian writes:

> "Our human, social, and civic dilemmas are becoming technical. And our technical dilemmas are becoming human, social, and civic. Our successes and failures alike in getting these systems to do 'what we want,' it turns out, offer us an unflinching, revelatory mirror."[2]

Thus, the emergent danger comes from the effectiveness and the propensity of such systems to replicate, reinforce, or amplify harmful existing discriminatory acts.[3]

The following report is divided into three sections. The first, *How We Got Here*, focuses on the history and key concepts behind algorithmic bias, such as power and classification. It also provides a brief overview of the elements that constitute algorithmic production—from data collection to its processing to the models that feed into machine learning algorithms. Section two, *Bias In The Wild,* offers real-world examples of discriminatory data practices in the housing and healthcare domains. The third and final section, *Where We Go From Here,* examines the opportunities from the technical, legal, and policy space to improve non-discriminatory practices in algorithmic bias. The hope is that we learn from our history and work to produce more inclusive and just data practices.

---

[1] Offered by Noble, S. U. as a key goal in reimagining how information online can and should function; from (2018). *Algorithms of Oppression.* New York University Press.
[2] Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values.* W.W. Norton & Company, Inc.
[3] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 33-44. https://doi.org/10.1145/3351095.3372873

We thank the Mastercard Center for Inclusive Growth for their support and thought-leadership in this space. We also thank data.org for their thoughtful feedback and collaboration throughout this exploration. Finally, we owe a huge debt of gratitude to the many different academics, journalists, civil society organizations, government researchers, activists, public policymakers, technologists, and legal scholars for their resolve to untangling the messy, complex nature of our human pursuits and its resulting technical artifacts.

Section 1. How We Got Here

> "The notion that truths, values, & culture can be codified, quantified, & extracted from how individuals behave under artificial conditions assumes that these amorphous concepts are external to lived experiences or social context."
>
> - Mona Sloane & Emanuel Moss (2019)[4]

Algorithmic bias is – at the core – about power. It is the literal codification of what knowledge, which values, and whose truth is more valued. It calls into question: who has the authority and privilege to designate pieces of knowledge; whose expertise and narrative is left out; and to what extent are these human decisions embedded across various technical systems? While current discussions about "algorithmic bias" have recently breached the public consciousness, we begin by acknowledging a long legacy of scholarship dedicated to examining the relationship between technology and society. History underscores that bias in technology systems is not a new phenomenon. Instead, we are at a moment in which we have both the tooling and critical capacity to recognize, analyze, and potentially address how automated decision-making systems have negatively affected specific communities. This section also provides a high-level overview of the evolution of machine learning techniques that are at the center of today's algorithmic bias critiques. It is in reviewing this full context—the history, theory, and technology—that we begin to see and appreciate the potency of the algorithmic bias critique. It is also within this full context that future solutions and interventions must be informed.

A brief history lesson.

In 1957, UK Prime Minister, Harold Macmillan, declared, "Let's be frank about it. Most of our people have never had it so good."[5] Emerging from a "bombed-out, financially and morally exhausted" period, the 1950s marked a significant social and cultural transformation for the country. The 1950s, now synonymous with Elvis and Alan Turing, also serves as one of the first prehistories of algorithmic bias. In her article, "Hacking the Cis-tem," historian Mar Hicks provides a thorough and powerful account of the ways in which technologies are political.[6] Specifically, Hicks uncovers a history of transgender Britons whose push to correct gender listings in government-issued ID cards during the mainframe era made visible "how systems were designed and programmed to accommodate certain people and to deny the existence of others."[7]

According to Hicks, the largest of the government's computer installations was housed within the Ministry of Pensions and National Security. The Ministry paid out pensions, kept track of the money paid into the welfare system by working Britons, and issued a National Insurance card which was required to

---

[4] Sloane, M., & Moss, E. (2019). AI's social sciences deficit. *Nat Mach Intell* 1, 330-331. https://doi.org/10.1038/s42256-019-0084-6

[5] Sandbrook, D. Fifties Britain: Never so good? Or too good to be true? The National Archives. https://www.nationalarchives.gov.uk/education/resources/fifties-britain/

[6] Hicks, M. (2019). Hacking the Cis-tem. *IEEE Annals of the History of Computing*, 41(1), 20-33. https://ieeexplore.ieee.org/document/8634814

[7] Hicks, M. (2019). Hacking the Cis-tem. *IEEE Annals of the History of Computing*, 41(1), 20-33. https://ieeexplore.ieee.org/document/8634814

claim pension benefits and or gain employment. To keep up with the massive recordkeeping requirements, the EMIDEC 2400 mainframe was installed. Hicks wrote:

> "The purpose of this system was to ensure the accurate and timely collection of taxes and disbursements of state pension payments. As a byproduct, this system—and others like it—took for granted particular identity categories and those designing the system programmed assumptions related to those identity categories into the computer, whose programing was built on pre-electronic rules for handling citizens' accounts."[8]

By the mid-1950s, hundreds of trans Britons had petitioned the Ministry of Pension to correct the gender on their benefits cards in order to gain legal employment. These requests for change, Hicks notes, resulted in a centralized list, making it possible to track every transgender citizen in the country. The list was shared, unbeknownst to the individuals, to government medical institutions for longitudinal medical studies on the nature of the trans identity and the process of transitioning.[9] To officially receive permission for a records change, described Hicks, required a certain level of resources and was often tied to ideals about class, sexual and gender normativity, and economic worthiness.[10] As the Ministry began to expand its use of computer systems, the procedures for amending the gender recorded also changed, requiring not just a change to one's birth certificate but for the record to be punched into the mainframe. Hicks wrote, "Despite the enhanced flexibility offered by computer methods, the digitization of people's accounts actually resulted in *less* flexibility and *less* accommodating policies for trans people."[11]

Additionally, the mechanism of record transfer—punching cards for the mainframe—serves as our first salient example of the way in which data plays a critical component of any systems function. "How information was categorized, sorted, and taxonomized before, during, and after being transferred to punched cards was itself a critical part of the system," noted Hicks. To this, the Ministry shifted course and no longer honored requests for gender changes by deliberately not altering the information on the data record. Instead, the Ministry described the new computing system as blind to gender. "In essence, the Ministry had decided to use the power of the new computer system to resubmerge trans citizens and their requests for recognition," wrote Hicks.[12]

Histories such as this serve as powerful reminders that the process of computerization and its impact on certain communities are socially, economically, and politically motivated. "The purpose of these systems is to discipline information in the service of a particular goal," wrote Hicks. "However, this process of

---

[8] Hicks, M. (2019). Hacking the Cis-tem. *IEEE Annals of the History of Computing*, 41(1), 20-33. https://ieeexplore.ieee.org/document/8634814
[9] Hicks, M. (2019). Hacking the Cis-tem. *IEEE Annals of the History of Computing*, 41(1), 20-33. https://ieeexplore.ieee.org/document/8634814
[10] Hicks, M. (2019). Hacking the Cis-tem. *IEEE Annals of the History of Computing*, 41(1), 20-33. https://ieeexplore.ieee.org/document/8634814
[11] Hicks, M. (2019). Hacking the Cis-tem. *IEEE Annals of the History of Computing*, 41(1), 20-33. https://ieeexplore.ieee.org/document/8634814
[12] Hicks, M. (2019). Hacking the Cis-tem. *IEEE Annals of the History of Computing*, 41(1), 20-33. https://ieeexplore.ieee.org/document/8634814

rendering information computable relies on institutionalizing the views and biases of those constructing the system and reflexively serves their ends."[13]

In theory.

In 1980, Langdon Winner published his seminal piece, "Do Artifacts Have Politics?" in which he introduced the theory of "technological politics." This theory, Winner wrote, offers a framework for interpreting and explaining the impact of large-scale sociotechnical systems. The theory adds to previous critiques of the social determination of technology by bringing our attention to the "characteristics of technical objects and the meaning of those characteristics."[14] Drawing from various real-world examples, such as the pneumatic molding machines introduced by McCormicks in 1880, to renewable energy, to the atom bomb, Winner explores not just the social factors driving the creation of these technologies, but questions the specific features in the design or arrangement of a device that establishes patterns of power and authority. Winner wrote:

> "The things we call 'technologies' are ways of building order in our world. Many technical devices and systems important in everyday life contain possibilities for many different ways of ordering human activity. Consciously or not, deliberately or inadvertently, societies choose structures for technologies that influence how people are going to work, communicate, travel, consume, and so forth over a very long time. In the process by which structuring decisions are made, different people are differently situated and possess unequal degrees of power as well as unequal levels of awareness."[15]

Fast forward to today, and we find numerous studies and writings that share, in part, Winner's original theory. Most notably, Safiya U. Noble's groundbreaking 2018 research, *Algorithms of Oppression: How Search Engines Reinforce Racism*, demonstrates how commercial engines like Google enable the act of "technological redlining." Through an analysis of media searches and online paid advertising, Noble exposes a process of "digital sense-making" that emphasizes the impact that classification and the organization of information has over certain communities, especially those who have no direct redress for how they are represented, categorized, and indexed through searches.

Noble also presents a problematic future in which "it will become increasingly difficult for technology companies to separate their systematic and inequitable employment practices, and the far-right ideological bents of some of their employees, from the products they make for the public."[16] Almost prophetic, Noble's warning has materialized in several instances across Big Tech in 2020 from advertising boycotts of Facebook[17] to Google's controversial firings of labor activists[18] and

---

[13]Hicks, M. (2019). Hacking the Cis-tem. *IEEE Annals of the History of Computing*, 41(1), 20-33. https://ieeexplore.ieee.org/document/8634814

[14] Winner, L. (1980). Do Artifacts Have Politics? *Daedalus,* 109(1), 123. http://www.jstor.org/stable/20024652

[15] Winner, L. (1980). Do Artifacts Have Politics? *Daedalus,* 109(1), 123. http://www.jstor.org/stable/20024652

[16] Noble, S. U. (2018). *Algorithms of Oppression.* New York University Press.

[17] Levy, S. (2020, August 6). *Facebook has more to learn from Ad Boycott*. Wired. https://www.wired.com/story/rashad-robinson-facebook-ad-boycott/

[18] Paul, K. (2020, December 2). *Google broke US law by firing workers behind protests*. The Guardian. https://www.theguardian.com/technology/2020/dec/02/google-labor-laws-nlrb-surveillance-worker-firing

prominent Black researchers[19] and to reports of rampant discrimination and misogyny at Pinterest.[20] But, *what do these examples have to do with bias in algorithms?*

The short answer is that bias, whether economic, political, or cognitive, influences how we interact with technology every day. It is also then reinforced by the technology itself. As Nicol Turner Lee of the Brookings Institute, described: "It is a virtuous cycle."

To get a handle on the various components, we break down this cycle to review concepts such as data, representation, classification, and organization—all of which inform and create the datasets used to train algorithmic decision-making systems. We begin with the five main data science stages: *Define and Collect; Label and Transform; Analysis and Insight; Modeling; and Deployment,* which are all informed by the data science field.[21] For our purposes, we focus on the first three stages and then map these onto concepts such as representation and classification.

1. *Define and Collect*

In data science, the process must begin with clearly defining the problem, whether it be for business or research. A clear definition of the problem statement is key to developing metrics for success and specific machine learning tasks (more on this later). Once defined, organizations then begin to discover and collect internal or external sources of data. Data collection is a systematic approach that is grounded in solving for the problem statement. This first stage is crucial for any attempt towards responsible data use that is principled and moral, particularly as it relates to personal information.[22]

2. *Label and Transform*

As data is collected, organizations may choose to standardize and organize the data to fit its purpose. This includes integration of data from diverse sources and or a relabeling of the data in order to create one structural dataset. This process transforms and may manipulate the data. As such, organizations are encouraged to perform quality checks and remediation on the data before moving on to analysis and modeling.

3. *Analysis and Insight*

The third stage is the beginning of exploratory data analysis (EDA). This step allows data scientists and engineers to begin applying certain tools and techniques to examine the data in a meaningful way, including regression analysis, descriptive statistics, visualizations, etc. The hope is to identify patterns, investigate nuances, and formulate modeling strategies.

---

[19] Ghaffary, S. (2020, December 9). *The controversy behind a star Google AI researcher's departure*. Vox. https://www.vox.com/recode/2020/12/4/22153786/google-timnit-gebru-ethical-ai-jeff-dean-controversy-fired

[20] Brougher, F. (2020, August 11). *The Pinterest Paradox: Cupcakes and Toxicity*. Medium. https://medium.com/digital-diplomacy/the-pinterest-paradox-cupcakes-and-toxicity-57ed6bd76960

[21] This is informed by the "Data Science Value Chain" from data.org

[22] *The Global Data Responsibility Imperative*. (2019, October). Mastercard. https://www.mastercard.us/content/dam/mccom/en-us/documents/global-data-responsibility-whitepaper-customer-10232019.pdf

On representation and classification.

On the surface, the first three stages are straightforward and are foundational for any data science project. But, to understand algorithmic bias and its intersection with these first three stages, we need to peel back a few more layers. History again proves to be a useful lens. In 2012, Steve Lohr wrote, "THIS has been the crossover year for Big Data—as a concept, as a term, and, yes, as a marketing tool. Big Data has sprung from the confines of technology circles into the mainstream."[23] Lohr's column cogently presented the lure of big data and its promises for unlocking its value using machine learning techniques. "In theory, Big Data could improve decision-making in fields from business to medicine, allowing decisions to be based increasingly on data and analysis rather than intuition and experience,"[24] wrote Lohr. This position, wherein quantification equates to objective and authoritative judgement, continues to permeate much of the current discussions around predictive and behavioral analytics. The problem, as scholars like Kate Crawford and danah boyd flagged, is that:

> "On one hand, Big Data is seen as a powerful tool to address various societal ills, offering the potential of new insights into areas as diverse as cancer research, terrorism, and climate change. On the other, Big Data is seen as a troubling manifestation of Big Brother, enabling invasions of privacy, decreased civil freedoms, and increased state and corporate control."[25]

boyd and Crawford's extended critique in "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," remains a formative piece for its interrogation of the underlying assumptions and bias associated with the computational culture of big data. In particular, this period saw a rise of research that leveraged social media and web data to answer questions about human-social phenomena. Two prominent examples include research on Twitter and the Arab Spring in 2011[26] and Google Flu Trends in 2014.[27] What we've come to learn since the earliest days of web-data driven analysis is that methods matter, a lot.

This leads us to the question of *why representation* matters in safeguarding against bias. We present two interpretations. First, statistical bias is considered to be the difference between an estimator's expected value and the true value given the data available. In statistics, this can be mitigated through a variety of methods, including the use of a representative sample. A sampling method is considered biased if it "systematically favors some outcomes over others."[28] This harkens back to the principle that data quality is paramount for any statistical analysis. Quality data will present features such as completeness and comprehensiveness, accuracy and precision, timeliness and relevance, etc. Rhetorically, then how representative is social media data or any user generated data online? Zeynep

---

[23] Lohr, S. (2012, August 11). *How Big Data Became So Big*. The New York Times. https://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html
[24] Lohr, S. (2012, August 11). *How Big Data Became So Big*. The New York Times. https://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html
[25] boyd, d., & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication, & Society*, 15(5), 662-679.
[26] Hounshell, B. (2011, June 20). *The Revolution Will Be Tweeted*. Foreign Policy. https://foreignpolicy.com/2011/06/20/the-revolution-will-be-tweeted/
[27] Lazer, D., & Kennedy, R. (2015, October 1). What we can learn from the epic failure of Google Flu Trends. Wired. https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/
[28] *Common Mistakes In Using Statistics: Spotting and Avoiding Them*. (n.d.). University of Texas at Austin. https://web.ma.utexas.edu/users/mks/statmistakes/biasedsampling.html

Tufecki points to four main methodological issues of social media big data studies: the prominent study of a single platform (e.g. Twitter); the selection of dependent variables (e.g. hashtags); an unrepresentative sample; and an omission of the wider social ecology of interaction.[29]

In another example, word embeddings, which are used in commercial search algorithms, center on a "word's representation" or numerical coordinates to establish similarity among other words. Christian explains that, "the embeddings, simple as they are—just a row of numbers for each word, based on predicting nearby missing words in a text—seemed to capture a *staggering* amount of real-world information."[30] With this technique, also came a realization of the gendered biases encoded in the English language. "For every clever or apt analogy for `man:woman`, like `fella:babe` or `prostate cancer:ovarian cancer`, there was a host of others that seemed to be reflecting mere stereotypes, like `carpentry:sewing` or `doctor:nurse`."[31] In machine learning systems, inferences from the data that produce such representations are not automatically recognized as biased. The system is, in fact, doing its job precisely as it was programmed to do—to identify the highest corollaries within a given data set. It is, therefore, not statistically biased.

But, satisfying the mathematical correctness and neutrality for an unbiased system is simply not enough. The real challenge, as Arvind Narayanan, posed is "how do we make algorithmic systems support human values?"[32] With this reframing, we can begin to recognize the complexity and the critiques of algorithmic systems to include issues around societal bias. This brings us to the second interpretation of representation that moves beyond data quality to also encompass the increasing need of varied perspectives and lived experiences in the design and development of algorithmic decision-making systems. As Virginia Eubanks writes in her book *Automating Inequality,* "We all inhabit this new regime of digital data, but we don't all experience it in the same way. . . Today, I mostly hear that the new regime of data constricts poor and working-class people's opportunities, demobilizes their political organizing, limits their movements, and undercuts their human rights."[33] Moreover, representation within the tech workforce has long been problematic and often construed as a pipeline problem with little progress.[34] Yet, recent data suggests that the pipeline is only part of the larger problem that reflects a culture of bias against the hiring and retention of minority populations. We explore this further in the next section which highlights specific examples.

Following the collection and evaluation of the data, comes another substantial stage in its preparation—the moment it is *labeled, classified and transformed*. One unique characteristic of big data analysis is the potential to stitch together various data sources into one large dataset. Does this

---

[29] Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*.
[30] Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values.* W.W. Norton & Company, Inc.
[31] Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values.* W.W. Norton & Company, Inc.
[32] Narayanan, A. [Arvind Narayanan]. (2018, March 1). *Tutorial: 21 fairness definitions and their politics* [video]. YouTube. https://www.youtube.com/watch?v=jIXIuYdnyyk&ab_channel=ArvindNarayanan
[33] Eubanks, V. (2017). *Automating Inequality*. St. Martin's Press.
[34] Harrison, S. (2019). "Five Years of Tech Diversity Reports—and Little Progress." *Wired.* https://www.wired.com/story/five-years-tech-diversity-reports-little-progress/

address the sample size disparity posed above? Not quite. Instead, implicit in this process of relabeling and reclassification is yet another set of human decisions that impact interpretability. As Tarleton Gillespie notes, "Just as one can know something about sculptures from studying their inverted molds, algorithms can be understood by looking closely at how information must be oriented to face them, how it is made *algorithm ready.*"[35] In the last several decades, information architecture has evolved from a more stringent set of hierarchical schemas to ones that now allow for more flexibility, such as relational and object-oriented databases as well as knowledge graphs. Inherent in any of this organization, however, remains the practice of categorization or classification. Any classification system embodies a dynamic compromise that encapsulates what belongs and what does not.[36] This negotiation is rarely made explicit or visible, and, in some instances, if done correctly fade into the background. Bowker and Star suggest that the question is not whether invisible organizational structures influence technology systems but rather how can we better recognize, learn from, and plan for their inevitable presence.[37]

By 2013, the discourse of big data and its accompanying technology tools reflected increasing concern over its political and social implications. Commercial and state sanctioned data surveillance programs as well as issues over online privacy served as key drivers to this debate. Unfortunately, as is the case today, we lack the standards and norms around how to appropriately balance the use of data-driven, algorithmic decision making equally and equitably. Solon and Selbst write,

> "Though useful, however, data is not a panacea. Where data is used predictively to assist decision making, it can affect the fortunes of whole classes of people in consistently unfavorable ways. Sorting and selecting for the best or most profitable candidates means generating a model with winners and losers. If data miners are not careful, the process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination."[38]

It is clear that we have not yet solved these concerns. Instead, critical examinations of big data have provided us the language necessary to engage in the discourse of algorithmic bias. In our reading, this is where we split hairs. Algorithmic bias, which builds off many of the concerns first expressed in the early days of the big data era, appends yet another layer of investigation—one that exposes and amplifies a systematic and repeatable error in computing systems that produces unfair outcomes.

---

[35] Gillespie, T. (2013). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski, K. A. & Foot (Eds.), *Media Technologies: Essays on Communication, Materiality, and Society* (pp. 167-193). MIT Press. https://www.intgovforum.org/multilingual/sites/default/files/webform/the_relevance_of_algorithms_-_tarleton_gillespie.pdf

[36] Bowker, G., & Star, S. (2000). *Sorting Things Out: Classification and its Consequences.* MIT Press.

[37] Bowker, G., & Star, S. (2000). *Sorting Things Out: Classification and its Consequences.* MIT Press.

[38] Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *104 California Law Review*, 671. https://ssrn.com/abstract=2477899

Of models and machine learning.

The technique known as "machine learning" is (loosely defined) as any algorithm that takes historical instances (aka training data) as its input to produce a decision rule or *classifier.*[39] The training data includes attributes of the data, called a feature, and the set of all available attributes defines the feature space or "representation" of the data. The classifier is then used on future instances of the problem. Hardt writes, "the whole appeal of machine learning is that we can infer absent attributes from those that are present" and that there is nothing that prevents the learning algorithm from not discovering these encodings.[40] The field of machine learning constitutes three major areas: *unsupervised learning* - a machine is given data and instructed to find patterns or make sense of it; *supervised learning* - a machine is given a series of categorized or labeled examples and told to make predictions about new examples; and *reinforcement learning* which places the system into an environment with rewards and punishments and is instructed to minimize punishment and maximize reward.[41] In unsupervised learning, we find model examples such as word-embedding, which builds vector word representations. Stated previously, research in this (specifically Google's word2vec and Stanford's GloVe models) space surfaced several encoded gendered stereotypes. In supervised learning, such as ImageNet, a set of predetermined categorizations assists the system to adjust the model for accuracy. These additional inputs are independent and identically distributed, thus having no direct causal connection to the output.[42] Reinforcement learning, on the other hand, generates a model that is trying to maximize for a reward and in which a previous decision influences the next decision.

Automated decision-making systems may leverage any one of the three machine learning approaches. In any instance, the use of a learning algorithm permits inferences and predictability, generating a sense of understanding and knowing derived from the dataset. As in statistics, it is in the confidence of the interpretation of these inferences and predictions that are often questioned. Furthermore, with reinforcement learning, changes within the model such as adjusting the weight of a parameter, may not be made explicit, generating a "black box" effect. The problem then becomes an inability to recognize and explain how certain variables are being combined for predictions. In 2020, Suresh and Guttag, released a framework for understanding the unintended consequences of machine learning. Figure 1 is helpful in visualizing the two phases of data generation (data science domain) and model building and implementation (machine learning domain).

[39] Hardt, M. (2014, September 26). *How big data is unfair.* Medium. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de
[40] Hardt, M. (2014, September 26). *How big data is unfair.* Medium. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de
[41] Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values.* W.W. Norton & Company, Inc.
[42] Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values.* W.W. Norton & Company, Inc.

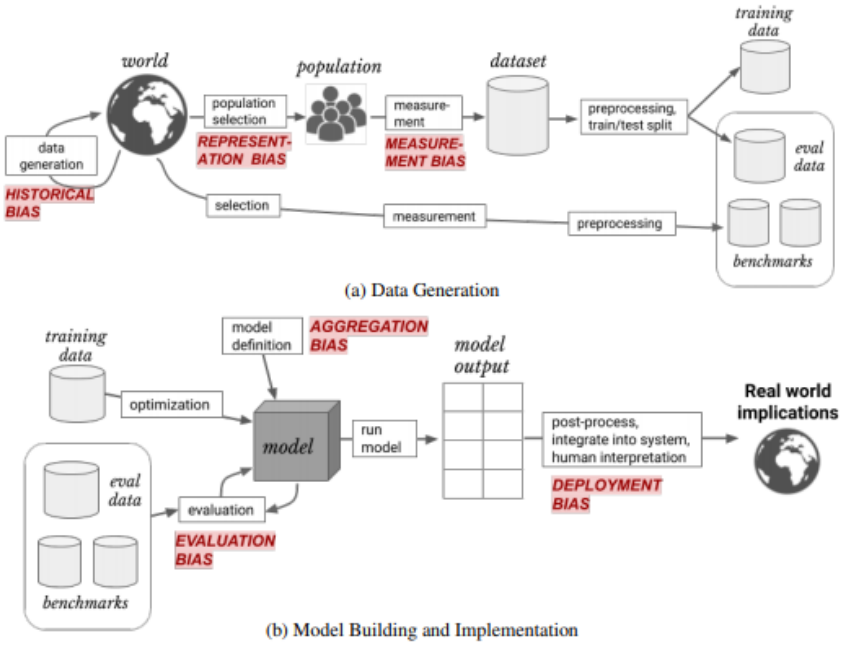

Figure 1: (a) The data generation process begins with data collection from the world. This process involves both sampling from a population and identifying which features and labels to use. This dataset is split into training and evaluation sets, which are used to develop and evaluate a particular model. Data is also collected (perhaps by a different process) into benchmark datasets. (b) Benchmark data is used to evaluate, compare, and motivate the development of better models. A final model then generates its output, which has some real world manifestation. This process is naturally cyclic, and decisions influenced by models affect the world that exists the next time data is collected or decisions are applied. In red, we indicate where in this pipeline different sources of downstream harm can arise.

Recognizing the sources for downstream harm, however, does not necessarily stop technology progress. In February 2015, DeepMind published a paper featuring a hybrid of classical reinforcement learning with neural networks, establishing what we now know as deep-learning.[43] We are only beginning to see glimpses into the consequential effects of deep-learning with deep fakes and manipulated media. And, while these seem currently insignificant and science-fiction-like, these examples are a testament to the sense of power imbued in these computing systems. This is only further complicated by a growing reliance - and blind faith - in such automated technologies to manage additional aspects of our daily lives more accurately, and more objectively, than any human might accomplish. Therefore, the more flexible and powerful learning systems become, the greater the need to understand what, exactly, they are learning to do on our behalf.[44] If alarms were ringing during the era of big data, then perhaps it's time to sound the klaxon.

[43] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, C., Legg, S., & Hassabis, C. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. https://doi.org/10.1038/nature14236

[44] Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values.* W.W. Norton & Company, Inc.

Section 2. Bias in the Wild

> "As we're on the cusp of using machine learning for rendering basically all kinds of consequential decisions about human beings in domains such as education, employment, advertising, health care, and policing, it is important to understand why *machine learning is not, by default, fair or just in any meaningful way.*"
>
> - Mortiz Hardt (2014)

Thanks to the work of researchers such as Cathy O'Neil, Helen Nissenbaum, Joy Buolomwini, Timnit Gebru, Frank Pasquale, Christian Sandvig, Latanya Sweeney, and many others, we have a growing catalogue of biases in technical systems. From healthcare, to online recommendation systems, to risk assessment tools, the ways in which we interact with automated decision-making systems is becoming more fluid and ubiquitous, exposing both differential and or discriminatory treatment. The following offers two well-documented examples of bias in the wild. The purpose of this section is to help bridge the theoretical to real-world examples. As with any abstract term, like an algorithm, articulating specific effects is useful in considering potential remedies and solutions. For our purposes, we triangulate around examples of marginalized populations and the effects of automated decision-making systems in housing and healthcare situations.

1. Housing

In 2019, the National Fair House Alliance (NFHA) released its annual report, "Defending against unprecedented attacks on fair housing," which brought to light concerns around new technologies used for advertising of homes and apartments, as well as for evaluating tenants and potential owners.[45] For instance, they found an automated system may reject a tenant's application based on residual correlations between race or ethnicity and wealth as well as other data sets.[46] In *Connecticut Fair Housing Center v. CoreLogic Rental Property Solutions*, a Connecticut federal district court ruled that a tenant screening company could be held liable for discrimination claims brought under the Fair Housing Act (FHA). Specifically, the court found CoreLogic's CrimSAFE product, which interprets an applicant's criminal record for tenant qualification, disqualified applicants for housing if the applicant was solely arrested but not convicted of a crime. The inferred corollary produced by CrimSAFE failed to account for racial disparities in arrest records.[47]

In the same year, NFHA in conjunction with other civil rights advocates announced a settlement with Facebook whose advertising platform was found to have enabled advertisers for housing the ability to exclude protected classes from receiving or viewing the housing advertisements. As a result, Facebook agreed to implement new procedures that would prevent advertisers from discriminating on the basis of the protected classes for ads related to housing, employment, and credit. This exclusion is further demonstrated by studies highlighting discriminatory lending practices in housing. A 2019 Berkeley study found that lenders using algorithms to generate decisions on loan pricing charged "otherwise-equivalent

---

[45] *Defending against unprecedented attacks on fair housing: 2019 Fair Housing Trends Report*. (2019). National Fair Housing Alliance. https://nationalfairhousing.org/wp-content/uploads/2019/10/2019-Trends-Report.pdf
[46] Sisson, P. (2019, December 17). *Housing discrimination goes high tech*. Curbed. https://archive.curbed.com/2019/12/17/21026311/mortgage-apartment-housing-algorithm-discrimination
[47] *Defending against unprecedented attacks on fair housing: 2019 Fair Housing Trends Report*. (2019). National Fair Housing Alliance. https://nationalfairhousing.org/wp-content/uploads/2019/10/2019-Trends-Report.pdf

Latinx/African-American borrowers 7.9 bps higher rates for purchase (refinance) mortgages, costing $765M yearly."[48]

The emerging technology issues in housing discrimination noted above have also produced, not surprisingly, non-technical ramifications. For, as we've argued throughout this piece, technology is both shaped by us and we are shaped by it. In September 2020, the Department of Housing and Urban Development (HUD) issued its final rule on the FHA disparate impact standard. "Under the disparate impact standards, employers, housing providers, and other entities are prohibited from using policies or practices that have a disproportionate negative impact on protected classes, even if they appear neutral on their face."[49] HUD's final rule removed the disparate impact test for most FHA cases, including those involving algorithmic tools. While these same tools have shown to create substantial risk of disparate impact discrimination, the elimination of the standard "establishes potentially insurmountable legal hurdles for victims of discrimination who seek to hold companies accountable for their reliance on biased algorithms." Unfortunately, under the Trump administration HUD's final rule not only scales back long-standing protections, it demonstrates, at minimum, a lack of understanding of the very real harms algorithmic systems can have on vulnerable communities.

2. Healthcare

The promise of technology innovation and the application of artificial intelligence has seen various success stories within the medical and healthcare fields. From determining a protein's 3D shape based on its amino-acid sequence, to virtual nursing assistants, to administrative workflow assistance, the healthcare industry has reaped the benefits of more efficient, more powerful automated systems. Like housing, however, healthcare also faces an enormous challenge in reckoning with large-scale issues of systemic bias. Specifically, recent research on a widely used algorithm for predicting risk — widely used by hospitals, health systems, insurance companies, and even government agencies — was found to exhibit significant racial bias.[50] The authors found that the "bias arises because the algorithm predicts health care costs rather than illness." They point out that Black patients generate fewer costs than White patients at the same level; however, this is largely due to unequal access and unequal treatment. Instead of using cost as the proxy, the authors suggest aligning for "high avoidable costs" and those who will have "high burden of active chronic conditions."[51]

In another example, Stanford researchers found that most clinical applications of deep learning algorithms are trained on US patient data in only three geographic areas, with cohorts disproportionately from

[48] Bartlett, R. P., Morse, A., Stanton, R. H., & Wallace, N. E. (2019, June). Consumer-Lending Discrimination in the Fintech Era. *National Bureau of Economic Research*, Working Paper 25943. https://www.nber.org/papers/w25943
[49] Sarkesian, L., & Sing, S. (2020, October 1). *HUDS new rule paves the way for rampant algorithmic discrimination in housing decisions*. New America. https://www.newamerica.org/oti/blog/huds-new-rule-paves-the-way-for-rampant-algorithmic-discrimination-in-housing-decisions/
[50] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019, October 1). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366, 447-453. https://escholarship.org/uc/item/6h92v832
[51] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019, November 1). *Algorithmic Bias in Health Care: A Path Forward*. Health Affairs. https://www.healthaffairs.org/do/10.1377/hblog20191031.373615/full/

California, Massachusetts, and New York.[52] This raises significant concerns about the validity of the algorithms for patients in other areas, introducing risk when implementing the diagnostic algorithms. "The data you have available impacts the problems you can study in the first place," commented Amit Kaushal, a co-author of the paper. "If I only have access to data from California, Massachusetts, and New York, I can build algorithms to help people in those places. But problems that are more common in other geographies won't even be on my radar."[53] Noted by the authors, this is not the first time medical data have lacked diversity. Early clinical trials often overlooked groups such as women, people of color, geographic locations, etc., which resulted in certain groups experiencing fewer benefits and more side effects from approved medications.[54]

The questions of what is the algorithm solving for and the data in its toolkit are foundational regardless of the domain. But, the devil is in the details, and deciding on which data should or should not be included is an ongoing conflict. In healthcare, a recent *New England Journal of Medicine* article suggests that "by embedding race into basic data and decisions in healthcare, these algorithms propagate race-based medicine."[55] The researchers included examples of race-adjusted algorithms from various practices such as cardiology to obstetrics. They go on to explain that "relationships between race and health reflect enmeshed social and biological pathways. . . Given this complexity, it is insufficient to translate a data signal into a race adjustment without determining what race might represent in the particular context."[56] Instead, the authors suggest an improvement to the algorithm would be to account for poverty, which is a significant predictor of other socioeconomic factors like housing, food insecurity, and life expectancy.

One final situational note. The COVID-19 pandemic will undoubtedly leave a lasting imprint on the livelihood of millions across the globe. At the time of this writing, an estimated 6.7 percent of Americans remain unemployed due to pandemic-related factors such as mandated stay-at-home orders.[57] Researchers also estimate that nearly 30–40 million people in America could be at risk of eviction in the next several months due to economic hardships as a result of COVID-19.[58] The pandemic has also brought to our attention a long overdue reckoning of healthcare disparities and social determinants of health among traditionally marginalized groups, which have disproportionately

---

[52] Kaushal, A., Altman, R., & Langlotz, C. (2020, September 22/29) Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA Network,* 324(12), 1212–1213. https://jamanetwork.com/journals/jama/article-abstract/2770833

[53] Lynch, S. (2020, September 21). *The Geographic Bias in Medical AI Tools*. Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/blog/geographic-bias-medical-ai-tools

[54] 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/

[55] Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020, August 27). Hidden in Plain Sight: Reconsidering the Use of Race Correction in Clinical Algorithms. *The New England Journal of Medicine*, 383(9), 874-882. https://www.nejm.org/doi/full/10.1056/NEJMms2004740

[56] Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020, August 27). Hidden in Plain Sight: Reconsidering the Use of Race Correction in Clinical Algorithms. *The New England Journal of Medicine*, 383(9), 874-882. https://www.nejm.org/doi/full/10.1056/NEJMms2004740

[57] Unemployment Rates During the COVID-19 Pandemic: In Brief. (2020, December). Congressional Research Service. https://fas.org/sgp/crs/misc/R46554.pdf

[58] Benfer, E., Robinson, D. B., Butler, S., Edmonds, L., Gilman, S., McKay, K. L., Neumann, Z., Owens, L., Steinkamp, N., & Yentel, D. (2020, August 7). *The COVID-19 Eviction Crisis: an Estimated 30-40 Million People in America Are at Risk*. The Aspen Institute Financial Securities Program. https://www.aspeninstitute.org/blog-posts/the-covid-19-eviction-crisis-an-estimated-30-40-million-people-in-america-are-at-risk/

been affected by COVID.[59] Moreover, as the global research community rushes to publish new findings, there is a risk of propagating biases or generating new biased models from the use of unrepresentative data samples and model overfitting.[60] Therefore, we must consider how today's social context can and will shape future datasets and models. While the pandemic is the common denominator across current issues in employment, housing, and healthcare, this should also serve as a reminder of how closely interconnected these basic needs can be, particularly for low-income persons. As such, we must take particular interest in untangling the rise of these "algorithmic webs,"[61] which leaves many trapped in the "digital poorhouse."[62]

[59] Facher, L. 9 (2020, May 19). *Ways Covid-19 may forever upend the U.S. healthcare industry*. Stat News. https://www.statnews.com/2020/05/19/9-ways-covid-19-forever-upend-health-care/
[60] Jerich, K. (2020, August 18). *AI bias may worsen COVID-19 health disparities for people of color*. Heathcare IT News. https://www.healthcareitnews.com/news/ai-bias-may-worsen-covid-19-health-disparities-people-color
[61] Hao, K. (2020, December 4). The coming war on the hidden algorithms that trap people in poverty. *MIT Technology Review.*
https://www.technologyreview.com/2020/12/04/1013068/algorithms-create-a-poverty-trap-lawyers-fight-back/
[62] Eubanks, V. (2017). Automating Inequality. St. Martin's Press.

Section 3. Where We Go From Here

"***algorithm*** (n.):
Word used by programmers when they do not want to explain what they did."
–unknown[63]

The first two sections of this report focus on the history, theory, and technology that has shaped how we have come to understand algorithmic bias in modern-day automated decision-making systems. It is clear that we find our own culture and society reflected in new technologies, including machine learning. For every concern raised by algorithmic bias, we are also compelled to find ways to neutralize it.

In this third and final section, we bring forth promising approaches for addressing bias in algorithms. Again, this is not a comprehensive list. Instead, the following highlights work that have garnered significant attention over the past few years. But, before we jump into specifics, it is important to recognize a set of practices that are generally useful and agnostic to domain and or a specific problem statement. These are derived from a litany of "best practices" found throughout data science and computer science communities: 1) Clearly define your problem; 2) Choose a learning model fit for the purpose; 3) Utilize a representative training data set; and 4) Monitor performance by simulating real-world conditions. These practices are particularly useful for those embarking on the start of building an algorithmic model. But, what of the algorithms already deployed? Are there strategies for retrofitting a pre-existing system against bias? What should operators of algorithms do to help prevent the introduction of any new biases into a system? The following offers several recommendations. Lastly, while each strategy finds its home in a specific discipline or community, neutralizing bias will require a mixture of several tactics. In other words, to mitigate algorithmic bias necessitates a multi-stakeholder, systems-level approach.

Ethical Frameworks.

Ethics is fundamental to any discussion of potential safeguards for machine learning and AI systems. Globally, there are over 160 AI ethics guidelines,[64] including the OECD AI Principles,[65] the EU's Ethics Guidelines for Trustworthy AI,[66] and Canada's Directive on Automated Decision-Making.[67] These guidelines are helpful in articulating high-level value statements from which various stakeholders and decision makers can incorporate into their own culture or product development cycle. While ethical frameworks and principles offer a roadmap, most are self-imposed, self-regulated, and difficult to implement, especially mathematically (e.g. fairness). And, in cases where AI assisted decision making technologies are deployed in social assistance systems, ethics are insufficient. Instead, there is significant and dire need for actionable interventions that may include third-party oversight and accountability

[63] Usman, M. (n.d.). *Programmers and Coders Wallpapers HD*. PCBots Lab. http://pcbots.blogspot.com/2013/07/coders-hd-wallpaper-by-pcbots.html
[64] AI Ethics Guidelines Global Inventory. (n.d.). https://inventory.algorithmwatch.org/
[65] OECD Principles on AI. (n.d.). OECD. https://www.oecd.org/going-digital/ai/principles/
[66] High-Level Expert Group on AI. (2019, April 8). *Ethics Guidelines for Trustworthy Artificial Intelligence*. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[67] Directive on Automated Decision-Making. (2019, February 5). Government of Canada. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

measures. Some legal scholars and researchers are calling for AI to also be "[human] rights respecting."[68] We address this in further detail below.

Fairness, Accountability, and Transparency.

The terms "fairness," "accountability," and "transparency" each come with their own histories. For our purposes, however, we focus on the use of these terms within the machine learning community. In 2014, Borocas and Hardt introduced the first computing workshop on the concepts of fairness, accountability, and transparency at the Neural Information Processing Systems (NeurIPS) conference. In her opening session, renowned Harvard computer scientist Cynthia Dwork posed the question "can we learn to be fair?," which argued "that a classification is fair only when individuals who are similar with respect to the classification task at hand are treated similarly, and this in turn requires understanding of sub-cultures of the population."[69] The workshop, alongside similar efforts on recommender systems (FAT/REC), natural language processing (Ethics in NLP), and data and algorithmic transparency (DAT), now comprise the official ACM peer-reviewed conference (ACM FAccT).[70] Since its inception, the community has blossomed to include computer scientists, statisticians, social scientists, scholars of law, and others.

The work produced from this community has led to significant rethinking of what fairness," "accountability," and "transparency" mean in a digital context. For example, fair machine learning, which is motivated by the outcome that decisions guided by algorithms are equitable, wrestles with definitional boundaries as well as statistical limitations. Not surprisingly, there is no single definition of fairness within computer science or statistics.[71] The definitions can change relative to the context and model at play. However, Davies-Corbett and Goel have identified three formal definitions: calibration (e.g. conditional on risk estimates), anti-classification parity (e.g. protected attributes, like race and gender) or classification parity (e.g. common measures of predictive performance),[72] which have been found to "perversely harm the very groups they were designed to protect."[73] Such findings have had rippled effects into other domains. In a 2017 *University of Pennsylvania Law Review* article, the authors note that "there may be cases where allowing an algorithm to consider protected class status can actually make outcomes

[68] Hilligoss, H., Filippo, A., Raso, F.A., & Krishnamurthy, V. (2018, October) *It's not enough for AI to be 'ethical'; it must also be 'rights respecting.'* Berkman Klein Center Medium. https://medium.com/berkman-klein-center/its-not-enough-for-ai-to-be-ethical-it-must-also-be-rights-respecting-b87f7e215b97

[69] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2014). *Can we learn to be fair?* Fairness, Accountability, and Transparency in Machine Learning, Montréal, Canada. https://www.fatml.org/schedule/2014/presentation/can-we-learn-be-fair-2014

[70] ACM FAT* Conference Examines Fairness, Accountability and Transparency of Algorithmic Systems. (2019, January) Association for Computing Machinery. https://www.acm.org/media-center/2019/january/fat-2019

[71] Narayanan, A. [Arvind Narayanan]. (2018, March 1). *Tutorial: 21 fairness definitions and their politics* [video]. YouTube. https://www.youtube.com/watch?v=jIXIuYdnyyk&ab_channel=ArvindNarayanan

[72] Hellman, D. (2019, July 11). *Measuring Algorithmic Fairness*. Virginia Public Law and Legal Theory Research Paper 2019-39. https://ssrn.com/abstract=3418528

[73] Corbett-Davies, S., & Goel, S. (2018, August 14). *The Measure and Mismeasure of Fairness: A Critical Review of Fair MAchine Learning*. arXiv. https://arxiv.org/abs/1808.00023

fairer" requiring a "doctrinal shift" in the consideration of protected status as presumptively a legal harm.[74]

Accountability and transparency have faced similar reflections. For example, the power and flexibility of neural net based systems are often critiqued as opaque and unintelligible, meaning even its creator may not be able to explain how the model came to its conclusion. This is deeply concerning as such techniques are proliferating across industry, government, military and medicine. Transparency is often recognized as one mitigation strategy and is equated with the need to see and observe a phenomenon in order to bring about accountability and governance. Transparency can be at differing levels such as platform design down to a system's source code. However, as Kroll et. al, observe: "source code alone teaches a reviewer very little, since the code only exposes the machine learning method used and not the data-driven decision rule"[75] Instead, Ananny and Crawford suggest reimagining transparency in an algorithmic system "not just as code and data but as an *assemblage* of human and non-human actors." In doing so, we must reconstruct both transparency and accountability to effect across a system, rather than inside a particular context.[76]

Finally, while the discussions around fairness, accountability, transparency, and ethics in machine learning are less than a decade old, the depth and breadth of proposed solutions is noteworthy. For example, there are now multiple toolkits, such as IBM's AI Fairness 360[77] (AIF360-open source Python toolkit for algorithmic fairness) or Amazon's SageMaker Clarify,[78] that help facilitate evaluations of algorithms, identify bias in datasets, and explain predictions. We also see growing interest in additional efforts on explainable AI, AI system factsheets,[79] datasheets for datasets,[80] bias impact statements,[81] and many others.

---

[74] Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, 165(3), 633-705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3

[75] Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, 165(3), 633-705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3

[76] Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989. https://doi.org/10.1177/1461444816676645

[77] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Majsilovic, A. Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. IBM Research. https://arxiv.org/pdf/1810.01943.pdf

[78] Amazon SageMaker Clarify. (n.d.). Amazon Web Services. https://aws.amazon.com/sagemaker/clarify/

[79] Richards, J., Piorkowski, D., Hind, M., Houde, S., & Mojsilović, A. (2020, June). *A Methodology for Creating AI FactSheets*. arXiv. https://arxiv.org/abs/2006.13796

[80] Gebru, T., Morgenstern, J., Vecchione, B., Vaughn, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). *Datasheets for Datasets*. Fairness, Accountability, and Transparency in Machine Learning. https://www.fatml.org/media/documents/datasheets_for_datasets.pdf

[81] Turner, N. L., Resnick, P., & Barton, G. (n.d.). *Algorithmic bias detection and mitigation: best practices and policies to reduce consumer harms*. Brookings Institute. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Algorithmic Auditing.

An algorithmic audit is another strategy to expose systematic biases embedded in software platforms. These audits serve as a bridge between the concerns raised in ethics discussions and tangible, actionable solutions. Christian Sandvig et. al, point to the use of "audit studies," which are commonly used in social science research, to ascertain whether algorithms result in harmful discrimination.[82] A number of algorithm audits have been conducted over the years, including detecting unfairness in online advertising and online search and price discrimination on e-commerce websites.[83]

In their 2019 paper on publicly naming biased performance results of commercial AI products, Raji and Buolamwini present an audit design and disclosure procedure that engages companies to reevaluate and make model improvements to address classification biases in their systems.[84] This approach, informed by the information security practice of "coordinated vulnerability disclosures" (CVD) and bug bounties, requires a neutral third-party to conduct the audit. The critical lever, however, is the public disclosure of these performance vulnerabilities which applied an external pressure necessary to enact change by the targeted companies.

Recently, Google and the Partnership for AI created and published a framework for auditing AI systems targeted at engineering teams. The framework, named Scoping, Mapping, Artifact Collection, Testing, and Reflection (SMACTR), is captured in Figure 2. Unlike previous audits processes, this work provides an end-to-end framework to be applied throughout the internal organization development life-cycle. The SMACTR audit is informed by other fields "where safety is critical to protect human life, such as aerospace and health care, which now carry out audits as part of the design process."[85]

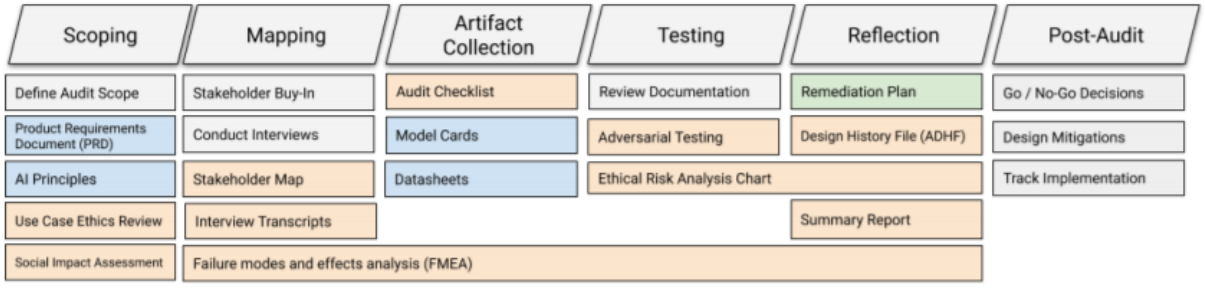

Figure 2: Overview of Internal Audit Framework. Gray indicates a process, and the colored sections represent documents. Documents in orange are produced by the auditors, blue documents are produced by the engineering and product teams and green outputs are jointly developed.

[82] Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). An Algorithm Audit. In Seeta Peña Gangadharan (Eds.), *Data and Discrimination: Collected Essays* (pp. 6-10). New America Foundation.
[83] Auditing Algorithms. (n.d.). Algorithm Audits by Researchers. https://auditingalgorithms.science/?p=153
[84] Raji, I., & Buolamwini, J. (2019). *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*. Conference on Artificial Intelligence, Ethics, and Society, Honolulu, HI, United States. https://dam-prod.media.mit.edu/x/2019/01/24/AIES-19_paper_223.pdf
[85] Johnson, K. (2020, January 30). *Google researches release audit framework to close AI accountability gap*. VentureBeat. https://venturebeat.com/2020/01/30/google-researchers-release-audit-framework-to-close-ai-accountability-gap/

Inclusive Design

Since the earliest commentaries on big data, one consistent critique has been the omission of certain populations from both representation in datasets and the design, deployment, and evaluation of automated decision-making systems. This omission can be the result of sampling methods, incomplete or unrepresentative training data, historical human biases, and a myriad of other reasons. As a best practice, developers and operators of algorithms should consider a more inclusive design approach. One specific tactical recommendation is to develop an algorithmic impact assessment tool which can help identify the appropriateness of certain automated decisions, and the severity of consequences for members of an affected group if not detected and mitigated.[86]

Another example of inclusive design comes from Google Research's People + AI Research (PAIR) initiative, which was first introduced in 2017. Central to PAIR's projects is the core idea of "participatory machine learning." This idea, like user-centric design, puts the human at the center of building AI technology. It is an attempt to empower and diversify stakeholders to think through what people need and how technology can help.[87] The group proposes three stages in which any community can and should participate: data training, defining success, and deployment. PAIR also released a guidebook[88] with resources for implementing machine learning projects that are responsive to user needs and ethical responsibility, as well as a more public policy oriented resource with non-technical explanations about machine learning.[89] The Mechanism Design for Social Good (MD4SG) is another example of a multi-institutional, interdisciplinary approach that includes collaborations and partnerships with impacted communities.[90] Founded in 2016, the group focuses on a range of domains to improve equity and social welfare for marginalized groups. Research produced by this group spans topics such as civic participation to data economies to climate change.

The benefits of inclusive design are both business and ethically oriented.[91] With a human-centric approach, products are better designed and built in response to specific user needs. It also serves as a check against assumptions about certain populations and as a barometer against potential output errors. The flip-side to inclusive, participatory design is that it takes time. Facilitating appropriate and nuanced engagement with certain communities, navigating differences, and formalizing all insights into product design may prove not to be efficient or scalable. Recognizing this is critical in deciding whether to pursue a participatory process.

---

[86] Turner, N. L., Resnick, P., & Barton, G. (n.d.). *Algorithmic bias detection and mitigation: best practices and policies to reduce consumer harms*. Brookings Institute. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

[87] Q&A: Participatory Machine Learning. (2020, July). People + AI Research Medium. https://medium.com/people-ai-research/participatory-machine-learning-69b77f1e5e23

[88] People + AI Guidebook. (2019, May 8). People + AI Research. https://pair.withgoogle.com/guidebook/

[89] AI Explorables. (n.d.). People + AI Research. https://pair.withgoogle.com/explorables/

[90] About. (n.d.). Mechanism Design for Social Good. http://md4sg.com/aboutus.html

[91] Trivedi, D. (n.d.). Insights on inclusive, human-centered AI: Meet PAIR co-founder Jess Holbrook. Accelerate with Google. https://accelerate.withgoogle.com/stories/insights-on-inclusive-human-centered-ai-meet-pair-cofounder-jess-holbrook

Public policy and regulation.

While much of the discussion above centers on the role technologists and operators of algorithms may play in mitigating bias, policymakers and regulators serve a critical function in this space. For example, the GDPR poses several actions that require a better formulation for explainability, transparency, and accountability of automated decision systems. In the U.S., several state-level regulators have placed moratoriums on the use of facial recognition software in response to rising civil liberty concerns. Noted above, in addition to establishing ethical principles, legal scholars are calling upon companies to follow an approach grounded in the Universal Declaration of Human Rights (UDHR). Human rights law provides a clear, legally binding strategy against discriminatory impacts.[92]

Alternative legal concepts, such as information fiduciaries and data trusts, are also taking shape. Positioned as "anticipatory regulation," this approach is considered to be more nimble, inclusive, collaborative and future-facing—all characteristics missing in today's regulatory environment.[93] In general, these legal concepts call for new data governance models, whether in the form of a trust, personal data stores, or digital fiduciaries. Regardless of which model, the goal is to rebalance power between the individual end user and those (e.g. companies) that collect, manage, and use our data. The realization of any of these emergent strategies requires both capacity and resources for additional research and piloting. One suggestion is to create an alliance to develop a Fiduciary Tech or "Fidtech" sector. FidTech would consist of companies and organizations that would be necessary to "offer protection and mediating support based on fiduciary principles, in balance to the increasing power of institutional AI.

[92] Raso, F. A., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018, September 26). *Artificial Intelligence & Human Rights: Opportunities & Risks*. Berkman Klein Center Research Publication 2018-6, https://ssrn.com/abstract=3259344

[93] *Data Trusts: A new tool for data governance*. (n.d.). ElementAI. https://hello.elementai.com/rs/024-OAQ-547/images/Data_Trusts_EN_201914.pdf

Conclusion

Bias in algorithms and AI systems is a complex issue. The solution is not just in adding more data sets, even if diverse. As illustrated in the previous sections, bias can and will be introduced at multiple points of a system's design and development. And as a result, algorithms can amplify systemic discrimination. However, this is not a reason to abandon the use of algorithms. Rather, it is a clear directive to recognize and address the issues raised—from incomplete and unrepresentative data; to a lack of diversity among decision makers, developers, and funders of AI tools; to a need for more inclusive design; and, for a re-evaluation of current legal and regulatory mechanisms.

Pointing back to Naraynan's original provocation at the beginning of this analysis, the problem is not purely about mathematical (or statistical) correctness. Instead, the challenge is how to make algorithmic systems support human values. As Christian suggests, perhaps it's an alignment problem.[94] Building on this, we put forward several topics for further discussion:

Interventions over predictions. Reframe the need for predictive analysis in risk assessment algorithms. As Zittrain et. al suggest, "machine learning should not be used for prediction, but rather to surface covariates that are fed into a causal model for understanding the social, structural and psychological drivers of crime."[95]

Inhibit the use of certain algorithmic systems. Similar to the discussions around facial recognition and its resulting moratoriums, there is a critical need to discuss which, if any, algorithmic systems should not exist and or be used in certain domains.

Develop standard for algorithmic impact assessment. Determining whether an algorithm is good or bad remains a key question for many policymakers. Drawing from the GDPR's data protection impact assessment (DPIA) obligation, a standard for algorithmic impact should be a necessary accountability component for any algorithm operator.

Reward for responsible algorithms. Celebrate companies that leverage several of the neutralizing tactics discussed above such as machine learning audits, inclusive design practices, representative sampling, diversified workforce etc.

Realigning internal employee metrics. Encourage CEO's to empower HR departments to formally incentivize employees to integrate DEI into work and development processes. In addition, organizations could help facilitate the exchange and deployment of technical expertise into community-based organizations in the form of a sabbatical or service hours.

---

[94] Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. W.W. Norton & Company, Inc.

[95] Zittrain, J. L., Barabas, C., Dinakar, K., Ito, J., Virza, M. (2018). "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment." Fairness, Accountability and Transparency in Machine Learning. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:62-76.

Levers in policymaking. While municipal and state-level governments have proven to be successful laboratories for experimentation of algorithmic-informed regulation, federal government agencies may also prove to be useful allies, allowing for additional tools such as rulemaking procedures.

Civil rights approach is insufficient. The traditional, legal approach of civil rights is insufficient to address the disparities enabled by automated decisions making systems, overlooking socioeconomic conditions. Therefore, a shift towards a human-rights approach may be warranted.

In closing, the purpose of this analysis is to provide a high-level overview of key topics and issues driving the study of algorithmic bias in automated decision-making systems. With an increasing reliance on such systems, today's research and critiques will be formative in how we move towards a more fair and equitable digital future.

References

About. (n.d.). Mechanism Design for Social Good. http://md4sg.com/aboutus.html

ACM FAT* Conference Examines Fairness, Accountability and Transparency of Algorithmic Systems. (2019, January) Association for Computing Machinery. https://www.acm.org/media-center/2019/january/fat-2019

AI Ethics Guidelines Global Inventory. (n.d.). https://inventory.algorithmwatch.org/

AI Explorables. (n.d.). People + AI Research. https://pair.withgoogle.com/explorables/

Amazon SageMaker Clarify. (n.d.). Amazon Web Services. https://aws.amazon.com/sagemaker/clarify/

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989. https://doi.org/10.1177/1461444816676645

Auditing Algorithms. (n.d.). Algorithm Audits by Researchers. https://auditingalgorithms.science/?p=153

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. 104 California Law Review, 671. https://ssrn.com/abstract=2477899

Bartlett, R. P., Morse, A., Stanton, R. H., & Wallace, N. E. (2019, June). Consumer-Lending Discrimination in the Fintech Era. *National Bureau of Economic Research*, Working Paper 25943. https://www.nber.org/papers/w25943

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Majsilovic, A. Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. IBM Research. https://arxiv.org/pdf/1810.01943.pdf

Benfer, E., Robinson, D. B., Butler, S., Edmonds, L., Gilman, S., McKay, K. L., Neumann, Z., Owens, L., Steinkamp, N., & Yentel, D. (2020, August 7). *The COVID-19 Eviction Crisis: an Estimated 30-40 Million People in America Are at Risk*. The Aspen Institute Financial Securities Program. https://www.aspeninstitute.org/blog-posts/the-covid-19-eviction-crisis-an-estimated-30-40-million-people-in-america-are-at-risk/

Bowker, G., & Star, S. (2000). *Sorting Things Out: Classification and its Consequences*. MIT Press.

boyd, d., & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication, & Society*, 15(5), 662-679.

Brougher, F. (2020, August 11). *The Pinterest Paradox: Cupcakes and Toxicity*. Medium. https://medium.com/digital-diplomacy/the-pinterest-paradox-cupcakes-and-toxicity-57ed6bd76960

Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. W.W. Norton & Company, Inc.

Common Mistakes In Using Statistics: Spotting and Avoiding Them. (n.d.). University of Texas at Austin. https://web.ma.utexas.edu/users/mks/statmistakes/biasedsampling.html

Corbett-Davies, S., & Goel, S. (2018, August 14). *The Measure and Mismeasure of Fairness: A Critical Review of Fair MAchine Learning*. arXiv. https://arxiv.org/abs/1808.00023

Data Science Value Chain. (n.d.). data.org.

Data Trusts: A new tool for data governance. (n.d.). ElementAI. https://hello.elementai.com/rs/024-OAQ-547/images/Data_Trusts_EN_201914.pdf

Defending against unprecedented attacks on fair housing: 2019 Fair Housing Trends Report. (2019). National Fair Housing Alliance. https://nationalfairhousing.org/wp-content/uploads/2019/10/2019-Trends-Report.pdf

Directive on Automated Decision-Making. (2019, February 5). Government of Canada. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2014). *Can we learn to be fair?* Fairness, Accountability, and Transparency in Machine Learning, Montréal, Canada. https://www.fatml.org/schedule/2014/presentation/can-we-learn-be-fair-2014

Eubanks, V. (2017). Automating Inequality. St. Martin's Press.

Facher, L. 9 (2020, May 19). *Ways Covid-19 may forever upend the U.S. healthcare industry*. Stat News. https://www.statnews.com/2020/05/19/9-ways-covid-19-forever-upend-health-care/

Gebru, T., Morgenstern, J., Vecchione, B., Vaughn, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). *Datasheets for Datasets*. Fairness, Accountability, and Transparency in Machine Learning. https://www.fatml.org/media/documents/datasheets_for_datasets.pdf

Ghaffary, S. (2020, December 9). *The controversy behind a star Google AI researcher's departure*. Vox. https://www.vox.com/recode/2020/12/4/22153786/google-timnit-gebru-ethical-ai-jeff-dean-controversy-fired

Gillespie, T. (2013). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski, K. A. & Foot (Eds.), *Media Technologies: Essays on Communication, Materiality, and Society* (pp. 167-193). MIT Press.

https://www.intgovforum.org/multilingual/sites/default/files/webform/the_relevance_of_algorithms_-_tarleton_gillespie.pdf

The Global Data Responsibility Imperative. (2019, October). Mastercard. https://www.mastercard.us/content/dam/mccom/en-us/documents/global-data-responsibility-whitepaper-customer-10232019.pdf

Hao, K. (2020, December 4). The coming war on the hidden algorithms that trap people in poverty. *MIT Technology Review.* https://www.technologyreview.com/2020/12/04/1013068/algorithms-create-a-poverty-trap-lawyers-fight-back/

Hardt, M. (2014, September 26). *How big data is unfair*. Medium. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

Harrison, S. (2019). "Five Years of Tech Diversity Reports—and Little Progress." *Wired.* https://www.wired.com/story/five-years-tech-diversity-reports-little-progress/

Hellman, D. (2019, July 11). *Measuring Algorithmic Fairness*. Virginia Public Law and Legal Theory Research Paper 2019-39. https://ssrn.com/abstract=3418528

Hicks, M. (2019). Hacking the Cis-tem. *IEEE Annals of the History of Computing*, 41(1), 20-33. https://ieeexplore.ieee.org/document/8634814

High-Level Expert Group on AI. (2019, April 8). *Ethics Guidelines for Trustworthy Artificial Intelligence*. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Hilligoss, H., Filippo, A., Raso, F.A., & Krishnamurthy, V. (2018, October) *It's not enough for AI to be 'ethical'; it must also be 'rights respecting.'* Berkman Klein Center Medium. https://medium.com/berkman-klein-center/its-not-enough-for-ai-to-be-ethical-it-must-also-be-rights-respecting-b87f7e215b97

Jerich, K. (2020, August 18). *AI bias may worsen COVID-19 health disparities for people of color*. Heathcare IT News. https://www.healthcareitnews.com/news/ai-bias-may-worsen-covid-19-health-disparities-people-color

Johnson, K. (2020, January 30). *Google researches release audit framework to close AI accountability gap*. VentureBeat. https://venturebeat.com/2020/01/30/google-researchers-release-audit-framework-to-close-ai-accountability-gap/

Hounshell, B. (2011, June 20). *The Revolution Will Be Tweeted*. Foreign Policy. https://foreignpolicy.com/2011/06/20/the-revolution-will-be-tweeted/

Kaushal, A., Altman, R., & Langlotz, C. (2020, September 22/29) Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA Network*, 324(12), 1212-1213. https://jamanetwork.com/journals/jama/article-abstract/2770833

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/

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, 165(3), 633-705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3

Lazer, D., & Kennedy, R. (2015, October 1). *What we can learn from the epic failure of Google Flu Trends*. Wired. https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Levy, S. (2020, August 6). *Facebook has more to learn from Ad Boycott*. Wired. https://www.wired.com/story/rashad-robinson-facebook-ad-boycott/

Lohr, S. (2012, August 11). *How Big Data Became So Big*. The New York Times. https://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html

Lynch, S. (2020, September 21). *The Geographic Bias in Medical AI Tools*. Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/blog/geographic-bias-medical-ai-tools

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, C., Legg, S., & Hassabis, C. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. https://doi.org/10.1038/nature14236

Narayanan, A. [Arvind Narayanan]. (2018, March 1). *Tutorial: 21 fairness definitions and their politics* [video]. YouTube. https://www.youtube.com/watch?v=jIXIuYdnyyk&ab_channel=ArvindNarayanan

Noble, S. U. (2018). *Algorithms of Oppression*. New York University Press.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019, October 1). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366, 447-453. https://escholarship.org/uc/item/6h92v832

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019, November 1). *Algorithmic Bias in Health Care: A Path Forward*. Health Affairs. https://www.healthaffairs.org/do/10.1377/hblog20191031.373615/full/

OECD Principles on AI. (n.d.). OECD. https://www.oecd.org/going-digital/ai/principles/

Paul, K. (2020, December 2). *Google broke US law by firing workers behind protests*. The Guardian. https://www.theguardian.com/technology/2020/dec/02/google-labor-laws-nlrb-surveillance-worker-firing

People + AI Guidebook. (2019, May 8). People + AI Research. https://pair.withgoogle.com/guidebook/

Q&A: Participatory Machine Learning. (2020, July). People + AI Research Medium. https://medium.com/people-ai-research/participatory-machine-learning-69b77f1e5e23

Raji, I., & Buolamwini, J. (2019). *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*. Conference on Artificial Intelligence, Ethics, and Society, Honolulu, HI, United States. https://dam-prod.media.mit.edu/x/2019/01/24/AIES-19_paper_223.pdf

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 33-44. https://doi.org/10.1145/3351095.3372873

Raso, F. A., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018, September 26). *Artificial Intelligence & Human Rights: Opportunities & Risks*. Berkman Klein Center Research Publication 2018-6, https://ssrn.com/abstract=3259344

Richards, J., Piorkowski, D., Hind, M., Houde, S., & Mojsilović, A. (2020, June). *A Methodology for Creating AI FactSheets*. arXiv. https://arxiv.org/abs/2006.13796

Sandbrook, D. *Fifties Britain: Never so good? Or too good to be true?* The National Archives. https://www.nationalarchives.gov.uk/education/resources/fifties-britain/

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). An Algorithm Audit. In Seeta Peña Gangadharan (Eds.), *Data and Discrimination: Collected Essays* (pp. 6-10). New America Foundation.

Sarkesian, L., & Sing, S. (2020, October 1). *HUDS new rule paves the way for rampant algorithmic discrimination in housing decisions*. New America. https://www.newamerica.org/oti/blog/huds-new-rule-paves-the-way-for-rampant-algorithmic-discrimination-in-housing-decisions/

Sisson, P. (2019, December 17). *Housing discrimination goes high tech*. Curbed. https://archive.curbed.com/2019/12/17/21026311/mortgage-apartment-housing-algorithm-discrimination

Sloane, M., & Moss, E. (2019). AI's social sciences deficit. *Nat Mach Intell* 1, 330-331. https://doi.org/10.1038/s42256-019-0084-6

Trivedi, D. (n.d.). Insights on inclusive, human-centered AI: Meet PAIR co-founder Jess Holbrook. Accelerate with Google. https://accelerate.withgoogle.com/stories/insights-on-inclusive-human-centered-ai-meet-pair-cofounder-jess-holbrook

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*.

Turner, N. L., Resnick, P., & Barton, G. (n.d.). *Algorithmic bias detection and mitigation: best practices and policies to reduce consumer harms*. Brookings Institute. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Unemployment Rates During the COVID-19 Pandemic: In Brief. (2020, December). Congressional Research Service. https://fas.org/sgp/crs/misc/R46554.pdf

Usman, M. (n.d.). *Programmers and Coders Wallpapers HD*. PCBots Lab. http://pcbots.blogspot.com/2013/07/coders-hd-wallpaper-by-pcbots.html

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020, August 27). Hidden in Plain Sight: Reconsidering the Use of Race Correction in Clinical Algorithms. *The New England Journal of Medicine*, 383(9), 874-882. https://www.nejm.org/doi/full/10.1056/NEJMms2004740

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, 109(1), 123. http://www.jstor.org/stable/20024652

Zittrain, J. L., Barabas, C., Dinakar, K., Ito, J., Virza, M. (2018). "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment." Fairness, Accountability and Transparency in Machine Learning. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:62-76.

Appendix

*Repository of organizations and projects related to algorithmic bias** (continually being updated)*

Type [1 - academic; 2 - industry; 3 - non-profit; 4 - government; 5 - philanthropy; 6 - other]	Name	URL	Key Topics
1	Berkman Klein Center at Harvard	https://cyber.harvard.edu/	AI Ethics, Algorithmic Justice; Data governance; Trustworthy AI
3	data.org	data.org	Data Science for Good; data and social impact
3	DataKind	https://www.datakind.org/	Data science for humanity
2	IBM AI Research	https://www.research.ibm.com/artificial-intelligence/publications/	Trustin AI, Fairness; Robustness; Explainability; Transparency and Accountability; Value Alignment
2	Google People + AI (PAIR)	https://pair.withgoogle.com/	Participatory Machine Learning
6	Association of Computing Machinery; FAT* Conference	https://facctconference.org/	Fairness, Accountability, Transparency
6	IEEE P7003 - Algorithmic Bias Working Group	https://sagroups.ieee.org/7003/	Benchmarking, validation, quality control; mitigation
3	Brookings Institute-Center for Technology	https://www.brookings.edu/center/center-for-technology-innovation/	Mitigation strategies and consumer protection
3	Human Rights Watch - Technology & Rights	https://www.hrw.org/topic/technology-and-rights	Personal data; privacy; surveillance; AI governance
6	OpenAI	https://openai.com/projects/	neural nets, language models, training scales
6	Partnership on AI	https://www.partnershiponai.org/	demographic data; responsible AI; facial recognition; benchmarking,

			explainableAI ; legal compatibility
2	Microsoft Research	http://www.jennwv.com/papers/checklists.pdf	Checklist to understand fairness in AI
1	UCLA Center for Critical Internet Inquiry	https://www.c2i2.ucla.edu/	interdisciplinary research; racial and economic justice
1	Internet Policy Research Initiative - MIT	https://internetpolicy.mit.edu/	Decentalization, AI policy; neural networks
3	data & society	https://datasociety.net/	Race and equity
3	Algorithmic Justice League	https://www.ajl.org/	Equitable and Accountable AI
1	Oxford Internet Institute	https://www.oii.ox.ac.uk/blog/tag/algorithmic-bias/	Governance, information inequality; digital knowledge and culture
1	Stanford Human-AI Center	https://hai.stanford.edu/research	Equitable and trustworthy AI systems
1	MIT Media Lab	https://www.media.mit.edu/research/?filter=groups	Human dynamics; social machines
1	Center for Ethics, Society, and Computing (ESC) and the Michigan Institute for Data	https://esc.umich.edu/	Ethics and Computing; social, cultural, and political dimensions of digital technologies.
1	USF Center for Applied Data Ethics and fast.ai	https://www.fast.ai/about/	democratizing deep learning systems