Artificial Intelligence and the Courts: MATERIALS FOR JUDGES

Artificial Intelligence and Bias – An Evaluation

SEPTEMBER 2022

AAAS
AMERICAN ASSOCIATION FOR
THE ADVANCEMENT OF SCIENCE

Artificial Intelligence and the Courts – Materials for Judges

The American Association for the Advancement of Science (AAAS) is honored to have been entrusted, by the National Institute of Standards and Technology (NIST), with the task of developing educational materials on artificial intelligence (AI) for judges and courts.

AAAS therefore offers this compilation of educational materials for judges, covering a wide, yet appropriate, set of issues. (Please see the list below). AAAS' goal is to provide a set of user-friendly and accurate, yet readily comprehended, definitions, analyses and perspectives, on a variety of terms and topics with which the judiciary ought to become familiar.

The materials contained herein were developed by teams of scientific and legal experts who focused on a particular topic. The topics considered worthy of inclusion were selected based both on the mandate provided by NIST and guidance received by AAAS from an Advisory Committee composed of a large and diverse group of legal and AI experts. Drafts of the materials were subsequently submitted to Advisory Committee members, and outside expert "Reviewers," to obtain any suggestions for adjustments before each team of authors finalized their contribution (paper, podcast, annex, etc.).

It is not expected that courts will become experts regarding these sometimes complex or technical matters. Rather, this collection presents facts and overviews in a manner intended to make judges aware of key issues and to enable courts to find useful information contained herein, easily.

Finally, it is hoped that courts will appreciate certain innovative elements of this product, notably the inclusion of podcasts. These will provide courts with facts and analysis of important questions in a format that courts may find agreeable and, given the accompanying transcripts included, useful. AAAS thanks NIST for allowing a team of experts to undertake this forward-leaning approach to providing courts with needed information and insights as part of this project.

Materials in this series include:

1. ***Artificial Intelligence – Foundational Issues and Glossary***
2. ***Artificial Intelligence and the Justice System*** **(Podcast Series and Transcripts)**

   Episode 1: *AI and Risk Scores* (49 minutes)

   Episode 2: *AI in the Legal Field – Commercial and Unexpected Uses* (70 minutes)

   Episode 3: *AI, Decision-Making, and the Role of Judges* (58 minutes)
3. ***Artificial Intelligence, Trustworthiness, and Litigation***
4. ***Artificial Intelligence, Legal Research, and Judicial Analytics***
5. ***Artificial Intelligence and Bias – An Evaluation***

Acknowledgements

The American Association for the Advancement of Science (AAAS) gratefully acknowledges the authorship of Michael Karanicolas, Executive Director of the Institute for Technology Law and Policy at the University of California, Los Angeles, and Mallory Knodel, Chief Technology Officer at the Center for Democracy and Technology.

We also extend our gratitude to the members of the Advisory Committee for their input and guidance throughout the design and production of this product: Rediet Abebe, Assistant Professor, Computer Sciences, University of California, Berkeley; Theodore F. Claypoole, Partner, Womble, Bond, Dickenson, LLP; John Cooke, Director, Federal Judicial Center; Cynthia Cwik, Senior Legal Advisor, Former Stanford Fellow, and Former Partner, Jones Day and Latham & Watkins; Herbert B. Dixon, Jr., Judge (ret.), Superior Court of the District of Columbia, and Senior Legal Advisor to the Center for Legal and Court Technology, William & Mary School of Law; Finale Doshi-Velez, Associate Professor, Computer Science, School of Engineering and Applied Sciences, Harvard University; Paul W. Grimm, Judge, US District Court for the District of Maryland; James Harris, Technology Consultant, National Center for State Courts; Joseph Sawyer, Director, Faculty Development and Online Learning, National Judicial College; Elham Tabassi, Chief of Staff, Information Technology Laboratory, National Institute of Standards and Technology; and Zachary Zarnow, Principal Court Management Consultant, National Center for State Courts.

AAAS appreciates the contributions of Solomon Assefa, Vice President, IBM Research; Nicole Clark, Chief Executive Officer, Trellis Research; Rebecca Crootof, Associate Professor of Law, University of Richmond School of Law; Patrick Huston, Brigadier General and Assistant Judge Advocate General for Military Law and Operations, US Department of Defense; Maia Levy-Daniel, Attorney and Public Policy Consultant; Matthew Stepka, Managing Partner, Machina Ventures, and Lecturer, Haas School of Business, University of California, Berkeley; Peter Stone, Professor, Department of Computer Sciences, and the Founder and Director, Learning Agents Research Group, Artificial Intelligence Laboratory, University of Texas, Austin; Rebecca Wexler, Assistant Professor of Law and the Faculty Co-Director, Berkeley Center for Law & Technology, University of California, Berkeley; and others for their review and insightful comments on previous drafts of the *Artificial Intelligence and the Courts: Materials for Judges* series.

This paper was edited by Joel Ericsen, Program Associate, and Alain Norman, Senior Program Associate, Science and the Law, AAAS. Special thanks to Theresa Harris, Program Director, Scientific Responsibility, Human Rights and Law Program, and Nathaniel Weisenberg, Program Associate, AAAS, and to Barbara “Bebe” Holloway, University of Virginia, for their contributions.

AAAS Creative and Brand Manager, Paula Fry, created the cover design for this publication.

The production of this publication and other materials in the *Artificial Intelligence and the Courts: Materials for Judges* series was funded by the United States National Institute of Standards and Technology (NIST) through Award 60NANB21D031.

Disclaimer: The opinions, findings and conclusions or recommendations expressed in this publication do not necessarily reflect the views of the AAAS Board of Directors, its council and membership, or NIST.

Contact: AAAS welcomes comments and questions regarding its work. Please send information, suggestions and any comments to the AAAS Scientific Responsibility, Human Rights and Law Program at srhrl@aaas.org.

Cite as: Karanicolas, M. and Knodel, M. "Artificial Intelligence and Bias – An Evaluation." *Artificial Intelligence and the Courts: Materials for Judges*. American Association for the Advancement of Science (2022). Available at https://doi.org/10.1126/aaas.adf0788.

Abstract

As the authors of this paper underscore, "Judges should expect that problems related to bias are likely to manifest, in one form or another, across virtually every field where AI decision-making has become popularized." Indeed, the problem-set of myriad ways bias can inhere in AI is of increasing interest and, to address it, a variety of perspectives is required; this paper cogently incorporates many important current viewpoints.* Courts will want to consider when or how bias relates to discriminatory impacts (e.g., in employment or housing cases), as well as the various possible origins of biases in a given AI system (e.g., datasets, design, or deployment). Also, difficulties with the "explainability" of AI may complicate showing the requisite discriminatory intent, but the "FAccT" (fairness, accountability and transparency) framework may help courts, as well as developers of AI systems, to understand the extent or nature of possible biases.

[*See also: NIST Special Publication (NIST SP) – 1270, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," Mar. 2022, at: https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence.]

Table of Contents

Artificial Intelligence and Bias – An Evaluation

Authors:

Michael Karanicolas
Executive Director, Institute for Technology Law and Policy, University of California, Los Angeles

Mallory Knodel
Chief Technology Officer, Center for Democracy and Technology

1. Introduction

As a growing proportion of our lives is governed by **artificial intelligence (AI) systems**[1] in both the private and public sphere, questions related to their accuracy and fairness have become increasingly pressing. Concerns about **bias** may seem counter-intuitive, since proponents of AI often point to its ability to remove ordinary markers of human bias from decision-making and replace subjective assessments around, say, a person's trustworthiness or neediness, with mechanically generated values.[2] However, there is a volume of research that demonstrates that not only can AI systems introduce novel harms and discriminatory impacts, but that biased or discriminatory **algorithms** may be even more dangerous than human decision-makers because algorithms hide behind a veneer of neutrality.[3]

This section introduces the origins of bias in automated decision-making as well as its impacts, across both the public and private sectors. It ends by considering these impacts' engagement with key legal concepts and the state of legal scholarship in assessing these questions. Our aim is to provide readers with a general introduction to bias in the context of AI, how AI biases impact society as well as the legal field, how the current legal framework addresses AI bias and what considerations judges should bear in mind in dealing with questions related to AI bias. The section begins with an overview of whether and how the legal system already addresses bias.

2. Bias and the Legal System

In order to properly frame our understanding of bias in AI systems, it is important first to consider a few avenues by which our legal system engages with both implicit and explicit bias. On an individual level, both lawyers and judges are typically expected to avoid discriminatory or harassing conduct. The American Bar Association's *Model Rules of Professional Conduct* section

[1] Bolded red terms appear in the Glossary.

[2] *See, e.g.* Kimberly A. Houser, *Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decision-Making*, 22 STAN. TECH. L. REV. 290 (2019).

[3] *See, e.g.* SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); Ruha Benjamin, *Assessing risk, automating racism* 366 SCIENCE 421 (2019); Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671 (2020); Michael Karanicolas, *To Err is Human, to Audit Divine: A Critical Assessment of Canada's AI Directive*, 14 J OF PARL. & POL. L. 1 (2019).

on "Misconduct" prohibits "conduct that the lawyer knows or reasonably should know is harassment or discrimination on the basis of race, sex, religion, national origin, ethnicity, disability, age, sexual orientation, gender identity, marital status or socioeconomic status in conduct related to the practice of law".[4] Judges, for their part, are required to perform their duties without bias or prejudice, to refrain from manifesting bias or prejudice, and to "administer justice without respect to persons."[5]

Historically, the most common legal questions related to bias typically manifested around employment or housing discrimination.[6] More recently, the criminal justice system has been a major area of focus, including discriminatory conduct by police,[7] lawyers,[8] judges,[9] jurors,[10] witnesses,[11] and even court personnel.[12]

These different categories may be further subdivided to include both **conscious and unconscious bias**, otherwise known as **explicit and implicit bias**, with the latter now being widely accepted as having a broad and significant impact across a range of decision-making and other cognitive functions.[13] Critically, the existence of implicit biases, even powerful ones, does not mean that individuals will always act in biased ways, particularly since these biases may be **consciously overridden**.[14] Nonetheless, as the legal profession has come to recognize the impact of bias on decision-making and outcomes, it has led to an imperative to consider the impacts of **structural biases** rather than attempting to root out overtly prejudiced individuals.

While there is no unified doctrine which the legal system uses to address bias, there are a number of principles which are relevant towards considerations of bias. First and foremost, the Fourteenth Amendment to the US Constitution prohibits state governments from denying a

[4] MODEL RULES OF PROF'L CONDUCT, R. 8.4 cmt. 3 (2020).

[5] MODEL CODE OF JUDICIAL CONDUCT R. 2.3 (2020); 28 U.S.C. § 453 (2006).

[6] *See, e.g.,* Brown v. Board of Education of Topeka, 347 U.S. 483 (1954); *a landmark case on school integration, and* Buchanan v. Warley, 245 U.S. 60 (1917), *which invalidated a city ordinance banning the sale of real property in particular neighborhoods to blacks.*

[7] Paul Butler, *Equal Protection and White Supremacy*, 112 NW. U. L. REV. 1457, 1461- 62 (2018).

[8] Irene Oritseweyinmi Joe, *Regulating Implicit Bias in the Federal Criminal Process*, 108 CALIFORNIA LAW REVIEW 965 970-974 (2020).

[9] Chris Guthrie, Jeffrey John Rachlinski, Sheri Lynn Johnson & Andrew J. Wistrich, *Does unconscious racial bias affect trial judges?,* 84 NOTRE DAME LAW REVIEW 1195 (2009).

[10] https://pubmed.ncbi.nlm.nih.gov/10508569/.

[11] John P. Rutledge, *They All Look Alike: The Inaccuracy of Cross-Racial Identifications*, 28 AM. J. CRIM. L. 207, 211-14 (2001).

[12] Debra Lyn Bassett, *Deconstruct and Superstruct: Examining Bias Across the Legal System*, 46 UC DAVIS LAW REVIEW 1563, 1579 (2013).

[13] *See, e.g.,* Anthony G. Greenwald et al., *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, 74 J. PERSONALITY & SOC. PSYCHOL. 1464 (1998), *which has been particularly influential in driving understandings of implicit bias.*

[14] Christine Jolls & Cass R. Sunstein, *The Law of Implicit Bias*, 94 CALIFORNIA LAW REVIEW 969, 974 (2006).

person within their jurisdiction the equal protection of its laws.[15] As a result of the Fifth Amendment, the same standards apply to the decisions of the federal government, such as prosecutorial decisions.[16] However, since 1976 the Supreme Court has required that plaintiffs show a discriminatory intent in order to establish a violation,[17] though this standard has been criticized for being outdated, given the volume of evidence related to the impacts of implicit bias.[18] Nonetheless, a law or policy which is neutral at face value will not be invalid under the Equal Protection Clause by virtue of having a more pronounced impact on one protected group than another.[19]

Direct intent is rare among AI systems, which generally return biased or discriminatory outcomes as a reflection of data or design flaws as opposed to overt instructions. As discussed in more detail in the following section, human bias can be introduced at each stage of development of AI systems. Together, these characteristics suggest that the Equal Protection Clause may not be a major source for developing case law in this space or, alternatively, that existing precedent is ill-suited to combat discrimination in an administrative context which is increasingly governed by AI.[20]

Where decisions emanate from administrative agencies, American law also requires there to be a "rational connection between facts and judgment."[21] While this standard accords significant deference to reviewing agency actions, it is potentially relevant to instances of bias or error among AI systems insofar as these decisions may fail to fulfill an adequate standard of transparency and **explainability**.[22]

A 2020 study revealed that nearly forty-five (45) percent of federal agencies have used either AI or machine learning for a range of functions, including enforcing regulatory mandates and adjudicating government benefits and privileges.[23] The nature of these systems makes them resistant to meaningful review of the rationale underlying particular decisions.[24] This suggests that American law would benefit from the development of new judicial standards to deal

[15] U.S. CONST. amend. XIV, § 1.

[16] Boiling v. Sharpe, 347 U.S. 497, 499 (1954).

[17] Washington v. Davis, 426 U.S. 229 (1976).

[18] *See, e.g.,* Yvonne Elosiebo, *Implicit Bias and Equal Protection: A Paradigm Shift*, 42 N.Y.U. REVIEW OF LAW & SOCIAL CHANGE 451 (2018), *which proposes a standard of discriminatory negligence for Equal Protection violations.*

[19] Washington v. Davis at 242.

[20] Yvonne Elosiebo, *Implicit Bias and Equal Protection: A Paradigm Shift*, 42 N.Y.U. REVIEW OF LAW & SOCIAL CHANGE 451 (2018).

[21] Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 56 (1983).

[22] Aram A. Gavoor, *The Impending Judicial Regulation of Artificial Intelligence in the Administrative State*, 97 NOTRE DAME LAW REVIEW REFLECTION 180, 184 (2022).

[23] DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO- FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 6–7 (2020), https://www-cdn.law.stanford.edu/wp- content/uploads/2020/02/ACUS-AI-Report.pdf.

[24] Pension Benefit Guar. Corp. v. LTV Corp., 496 U.S. 633, 654 (1990).

specifically with AI-based adjudications, and particularly with the unique challenges in developing robust due process protections in the context of relatively inscrutable outputs from an AI decision-maker.

Where AI decisions emanate from private sector agencies, such as banks, potential or current employers, biased decision-making could engage the Civil Rights Act,[25] the Americans with Disabilities Act (ADA),[26] and Section 503 of the Rehabilitation Act,[27] among others. However, Supreme Court precedent generally requires either "intent" or "motive" in discrimination for disparate treatment, which are difficult to ascribe in the context of an AI decisionmaker due to the fact that machines typically do not possess intentionality the way that humans do.[28]

Absent this "intent" or "motive", plaintiffs may still succeed by demonstrating that a practice disparately impacts a particular protected group.[29] If this disparate impact is sufficiently demonstrated, the burden shifts to the defendant as to whether the practice is "consistent with business necessity." If a practice is found to meet the standard of business necessity, the plaintiff can still prevail if they are able to demonstrate that a less discriminating but equally valid practice was available which the employer did not use.[30]

In carrying out an assessment of discrimination by a private sector entity, the lack of explainability underlying AI decisions can be a significant complicating factor, since it is difficult to pin down the underlying "practice" creating the disparate impact. A general decision to incorporate AI systems into the decision-making processes would be difficult, by itself, to establish such improper motive or intent, since many relatively benign factors, such as efficiency, might motivate a company to incorporate AI into its decision-making. Once again, the naturally human-centric manner in which jurisprudence has evolved, with its reliance on intent and rationale, runs into challenges in attempting to apply the same standards to an AI decision-maker.[31]

Finally, it is worth considering bias more structurally in the legal profession. Law is a distinctly human activity, which is grounded in relatively subjective assessments of concepts such as reasonableness, consent, or intent.[32] However, every individual involved in the administration of justice, from the policy-makers who draft the legislation, to lawyers and judges who interpret these concepts, approaches them through their own set of lived experiences, with all of the

[25] 42 U.S.C. § 2000e (2012).

[26] 42 U.S.C. § 12101 (2017).

[27] Rehabilitation Act of 1973, Pub. L. No. 93-112, 87 Stat. 355, 393 (codified as amended at 29 U.S.C. § 793).

[28] Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 335 n.15 (1977).

[29] Dothard v. Rawlinson, 433 U.S. 321, 329 (1977).

[30] 42 U.S.C. § 2000e-2(k) (2012); Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975).

[31] *For a more thorough discussion of this challenge in the employment context, see* Charles A. Sullivan, *Employing AI*, 63 VILLANOVA LAW REVIEW 395 (2018).

[32] Debra Lyn Bassett, *Deconstruct and Superstruct: Examining Bias Across the Legal System*, 46 UC DAVIS LAW REVIEW 1563, 1564 (2013).

baggage that can generate. None of us is truly a blank slate. Although there are a number of strategies for how bias can be confronted and mitigated, the growing recognition of the prevalence of bias has also been a key driver for diversity in the legal profession, in order to improve representation of perspectives and understandings of the law.[33] There are many novel aspects to how we think about bias in the context of AI, but at their core, these challenges are a continuation of a broader challenge to develop responsive and representative legal structures that reflect the spectrum of experiences and perspectives of the public they are meant to serve.

3. The Origins of Bias in AI Systems

The term "AI systems" is comprised of several technical concepts. AI is automation that aims to approximate human capability. Innovation in AI is often driven by the enthusiasm for exponentially increasing speed and scale of tasks through automation. Modern techniques to achieve automation include machine learning, deep learning and active learning.

Machine learning (ML) is a form of artificial intelligence algorithm that improves itself based on training data. The system "learns from experience." The way the machine "learns" depends on the algorithmic make-up of the system. **Deep learning** and active learning are more advanced techniques in which a system "learns how to learn" with (deep learning) or without (active learning) predetermined **datasets**.

Machine learning systems are enormous statistical inference engines with the capacity to generate outputs from the analysis of large inputs of data. Importantly, the data dependent nature of machine learning technology forms the basis of both the potentials and the pitfalls of contemporary artificial intelligence. Rather than eradicating human bias formed by the social and historical processes, such as racist, sexist, or ageist preconceptions, human bias leaks into AI technologies at every turn, exposing the very social tenets of what is commonly understood as purely technological, or rather technocratic. Machine learning bias, or what we refer to in this section as AI bias, then occurs when such algorithms produce outputs that are systemically prejudiced or discriminatory due to the underlying assumptions throughout various stages of the machine learning process.

There are many ways in which bias can find its way into AI: the structure of the data fed into the system and the architecture of the algorithm both have a valence for the biased outputs that such systems may generate. This is particularly problematic when such systems are employed to automate processes in social institutions, because if the bias in the system is not addressed, artificial intelligence technologies risk automating the inequalities inherent in our social systems.

[33] Sonia Sotomayor, *Lecture: 'A Latina Judge's Voice'*, N.Y. TIMES (May 14, 2009), https://www.nytimes.com/2009/05/15/us/politics/15judge.text.html.

The governance of AI is a question that we will return to when we look at AI deployment, a crucial and iterative final stage. A technology-centric approach to address the fairness, accountability and transparency issues in AI systems relies on a framework that breaks down the machine learning process into its constituent parts: design, development and deployment. We take these phases in turn as we uncover where bias originates in AI systems.[34]

3.1. Bias and AI Design

The structural and human biases present in society appear in the design of AI systems from the problem-solution generation stage and persist through the early-development stage. In her book *Race Against Technology*, scholar Ruha Benjamin exposes in great detail the ways "human decisions comprise the data and shape the design of algorithms, now hidden by the promise of neutrality and with the power to unjustly discriminate at a much larger scale than biased individuals."[35] For example, the problem is not only that predictive policing technologies are racially discriminatory, but that historically racialized groups are heavily policed and that predictive policing is seen as a way to scale up and automate the tasks required by over-policing communities of color.[36] In this way, bias has been introduced into an AI system before even one line of code or one data point has been collected.

Other examples where the design phase introduces bias include targeted advertising. While advertising merely aims to sell products to consumers, targeting is made possible through AI systems that are designed to take in data about individuals and code them as "interests." While targeted advertising may not explicitly aim to capture an individual's race, many targeted advertising systems nevertheless are able to code race as interests, through preferences for hair products, food, fashion or music to the degree that large advertising platforms promote their success at reaching audiences based on race. Other explicit captures of data to target by race include location and "ethnic affinity".

Another example that has a much greater potential for harm is when socio-economic status, including race, leads to inequalities in access to financial services, either because products are advertised differently based on perceived status or because the very determination of individuals' credit worthiness is made through the same data. It is clear to see how the design of targeted advertising of financial services might exacerbate the cycle of poverty, even if race and socio-economic status are not explicitly captured by advertising platforms.

[34] Vidushi Marda, '*Governance with Teeth*', ARTICLE 19 (April 2019), https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf.

[35] Ruha Benjamin, *Assessing Risk, Automating Racism: A health care algorithm reflects underlying racial bias in society,* 366 SCIENCE 6464, (2019).

[36] Where "racialized" refers to a socio-political process by which groups are ascribed a racial identity, whether or not members of the group self-identify as such; where "predictive policing" refers to technical tools and practices that use data and analytics to identify potential crimes; and where "over policing" refers to disproportionate police presence in marginalized communities that exacerbates poverty, marginalization and criminalization of individuals in those communities.

3.2. Bias in AI Development

There exist structural, statistical, socio-technical and human bias in the data, training procedures and validation stages of AI development. Each of these is taken in turn, below, to expose the origins of bias in AI systems.

3.3. Collecting Data and Data Sets

Machine learning requires vast amounts of data on which it must learn. This data is a major driver of bias in AI. Some sources of data are explicitly biased, such as troves of photographs and notations originating in eugenics research.[37] But in all cases, because "most machine-learning tasks are trained on large, **annotated data sets**… Such methods [of annotating training data] can unintentionally produce data that encode gender, ethnic and cultural biases."[38] While not all types of machine learning rely on predetermined data sets such as active learning, any potential model might therefore be built upon the broad **datification** of our deeply biased world.

For the most common types of AI that are trained on, and learn from, pre-determined data sets or data sources, the characteristics of the datasets used in machine learning fundamentally influence an AI model's behavior. A model is unlikely to perform well when it encounters novel data if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted societal biases.

"As a first step, researchers—across a range of disciplines, government departments and industry—need to start investigating how differences in communities' access to information, wealth and basic services shape the data that AI systems train on."[39] Scholars Kate Crawford and Ryan Calo are referring to how the data sets used in AI systems might themselves reproduce existing stereotypes by categorizing and inputting already differential manners in which various social groups access public resources.

The way in which data are collected from people also has weight in shaping the data sets. Here, institutional guidelines, as well as policies in tandem with documentational records, should be taken into account, as they have a direct effect on the types of data gathered. For example, NIST maintains a database of mugshot photos, in addition to other standard reference data, that is open and available as a training data set, yet it is under-representative of the US population in categories such as gender and race.[40] Mitigating potential bias in the use of such sets requires an open reckoning with such context.

[37] https://magazine.jhsph.edu/2022/how-biased-data-and-algorithms-can-harm-health

[38] James Zou, Londa Schiebinger, *Design AI so that it's fair,* 559 NATURE, 324-326 (2018).

[39] Kate Crawford, Ryan Calo, *There is a blind spot in AI research,* 538 NATURE, 311–313 (2016).

[40] "NIST Special Database 18." (2 December 2022). *NIST*. Available at: http://dx.doi.org/10.18434/T4159S.

3.4. Training AI Models

Another key aspect of AI development takes place in the training of an AI model. The AI model in its most basic form defines how input is turned into output, say for the purposes of natural language processing or computer vision. Training an AI model to perform according to the problem-solution generation of the design phase means active human supervision of the machine learning process as it takes in data and presents outputs. Tweaking **parameters**, **re-characterizing data**, using statistical methods and adjusting features of the model itself are ways that AI models are trained.

Human bias can be introduced at this stage because humans are doing some degree of supervising along a spectrum. Even in “unsupervised” machine learning, human feedback about the quality of the output is used by the algorithm to learn and optimize for future performance. Whether or not an answer is “right” can itself be a source of bias: For instance, an AI system might be tasked with sorting job applications to ensure new employees “fit in” in a company that has a problem with monoculture.

Some fixes are discussed by Zou and Schiebinger: “Thus, technical care and social awareness must be brought to the building of data sets for training. Specifically, steps should be taken to ensure that such data sets are diverse and do not under-represent particular groups. This means going beyond convenient classifications—‘woman/man’, ‘black/white’ and so on—which fail to capture the complexities of gender and ethnic identities.”[41] The more complex the AI system, the harder it is to avoid bias, say, when images classified as ‘nurse/doctor’ are perfectly paired with ‘woman/man.’[42]

3.5. Validating and Optimizing Outputs

As with the need to train or supervise machine learning, there must be an ongoing effort to validate outputs and optimize the model so as to increase accuracy, which some AI systems have been accused of lacking. Ensuing changes to the AI system can be made at any stage and take a variety of forms but are almost always determinations made by humans. Therefore, these determinations are points at which bias can be introduced. Adjusting the scope of the predetermined training data, data source, weights, parameters and other changes in the data or how the AI will assess data can confirm or further entrench the biases introduced at the design and development stages. Such adjusting does, however, also offer points at which de-biasing can be done.

It has been suggested that – at the dataset level—"every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended

[41] James Zou, Londa Schiebinger, *Design AI so that it’s fair,* 559 NATURE, 324-326 (2018).

[42] https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

uses."[43] This additional context and markers for datasets can increase **transparency** and **accountability** for developers. As the machine learning community at large has expressed a commitment to mitigate unwanted societal biases in machine learning models, it must work together to effectively reproduce machine learning results across diverse implementations. Another reason for increasing data set transparency and sharing results is so that researchers and practitioners can better select the appropriate datasets for their systems' goals.

At the training level, one technique to mitigate bias has been to statistically offset what is called "**word embeddings**" such as when nurse/doctor is equivalent to woman/man. Statistical offsets (i.e., consciously changing parameters) to avoid work embeddings lead to algorithms that "significantly reduce gender bias in embeddings, while preserving the useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias."[44]

Others take the validation and optimization phase as an opportunity to introduce auditing mechanisms such as, "an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups," published by Joy Buolamwini and Timnit Gebru.[45]

3.6. Bias in AI Deployment

Once a trained machine learning model has been integrated into an AI system its application to a problem-solution set requires mitigating the harms of the application of AI systems with fixes or improvements, or by lessening impacts. At the deployment stage, an AI system is often acting within an existing technocratic structure, for instance, to determine creditworthiness or the equitable delivery of social services. How an administrator or ultimate decision maker accounts for the AI system's output is a point at which bias can once again be introduced.

Virginia Eubanks' book "Automating Inequality" dramatically shows how data collected for technological purposes becomes a means of reinforcing economic marginality, which she refers to as "collective red-flagging, a feedback loop of injustice" (Eubanks, 2018:7). She criticizes the notion that a model is less biased than a human caseworker, homeless service provider, or intake caller; "I find the philosophy that sees human beings as unknowable black boxes and machines as transparent deeply troubling".[46]

[43] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortmanvaughan, Hanna Wallach, Hal Daumé III, Crawford Kate, *Datasheets for Datasets,* PROCEEDINGS OF THE 5TH WORKSHOP ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY IN MACHINE LEARNING, Stockholm Sweden (2018).

[44] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Sallgrama, Adam Kalai, *Manis to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,* 30TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS 2016), Barcelona Spain (2016).

[45] Buolamwini Joy, Gebru Timnit, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,* 81 PROCEEDINGS OF MACHINE LEARNING RESEARCH, 1–15 (2018).

[46] ibid:168

Eubanks poses two questions to assess the basic ethics of digital tools: (1) Does the tool increase the self-determination and agency of the poor? (2) Would the tool be tolerated if it was targeted at non-poor people?

These pointed questions can also be at odds with another set of questions about whether or not we "trust" the results of the AI system; whether or not the AI system's determination aligns with our own expectations. Caliskan writes, "Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names. Our methods hold promise for identifying and addressing sources of bias in culture, including technology."[47]

When models are built in one place, perhaps geographically, and then applied in another, unforeseen bias is much more likely to only be observable at the application phase. Shankar writes, "data sets appear to exhibit an observable amerocentric and eurocentric representation bias. Further, we analyze classifiers trained on these data sets to assess the impact of these training distributions and find strong differences in the relative performance on images from different locales. These results emphasize the need to ensure geo-representation when constructing data sets for use in the developing world."[48]

Yet there are additional questions beyond these pointed ethical considerations to more technological forms of AI governance such as the measurement of **fairness, accountability and transparency** (FAccT). The FAccT framework takes the middle ground between neutral and ethical technology to focus on questions of management: is it fair; is it accountable; is it transparent?

3.6.(a). The "FAccT" Framework

Engineers are often involved in designing the various auditing mechanisms that consider FAccT. Bellamy introduces an example of "a new open-source Python toolkit for algorithmic fairness, AI Fairness 360 (AIF360), released under an Apache v2.0 license (https://github.com/ibm/aif360). The main objectives of this toolkit are to help facilitate the transition of fairness research algorithms for use in an industrial setting and to provide a common framework for fairness researchers to share and evaluate algorithms."[49]

These mechanisms can be applied retroactively at every stage and need not wait until deployment. Friedler writes, "we find that **fairness-preserving algorithms** tend to be sensitive

[47] Caliskan Aylin, Bryson Joanna, Narayanan Arvind, *Semantics derived automatically from language corpora contain human-like biases,* 356 SCIENCE 6334, 183-186 (2017).

[48] Shankar Shreya, Halpern Yoni, Breck Eric, Atwood James, Wilson Jimbo, Sculley D, *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World,* Presented at NIPS 2017 WORKSHOP ON MACHINE LEARNING FOR THE DEVELOPING WORLD (2017).

[49] R.K.E. Bellamy, et al., *AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias*, 63 IBM JOURNAL OF RESEARCH AND DEVELOPMENT 4/5, 1-15 (2019).

to fluctuations in dataset composition (simulated in our benchmark by varying **training-test splits**) and to different forms of preprocessing, indicating that fairness interventions might be more brittle than previously thought."[50]

Yet there exists a final point along the spectrum from neutral to FAccT to ethical: Others have proposed a wider view of the social and human rights impacts of AI system deployment and application. Like Marda's work on AI governance cited above, Venkatsubramanian writes, "given that AI is no longer solely the domain of technologists but rather of society as a whole, we need tighter coupling of computer science and those disciplines that study society and societal values."[51] There is a now pervasive theory that altering the context in which AI engineering occurs, by innately considering human rights and society, will lead to more human rights-aligned outcomes.

Yet the widest view acknowledges that indeed AI systems are inherently embedded in the human world, and the human world is biased. Thus, "Even with careful review of the algorithms and data sets, it may not be possible to delete all unwanted bias, particularly because AI systems learn from historical data, which encodes historical biases."[52]

4. Impacts of AI Bias on Law and Society

The expanding footprint of algorithms in our day to day lives, otherwise known as the **algorithmic turn**, has led to a growing body of scholarship related specifically to concerns about fairness and bias.[53] From our daily search for news and information, to our choice of romantic partners, to our ability to find a job or a home, or to access credit, our lives and decisions are increasingly governed by invisible formulas designed to deliver efficiency, profit, engagement, or any number of other predetermined metrics.[54] The data-intensive networks that underlie these decisions, and the cold and impassive way in which the results are delivered, create an illusion of neutrality and fairness, especially as contrasted against the heavily subjective and instinct-driven processes that dominated most traditional decision-making.[55] However, as demonstrated by the previous sections, it is all too common for these algorithmic systems to

[50] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth, *A comparative study of fairness-enhancing interventions in machine learning,* PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT* '19), 329–338 (2019).

[51] Suresh Venkatasubramanian, Nadya Bliss, Helen Nissbaum, Melanie Moses, *Interdisciplinary Approaches to Understanding Artificial Intelligence's Impact on Society* (2020). arXiv:2012.06057

[52] Drew Roselli, Jeanne Matthews, Nisha Talagala, *Managing Bias in AI,* WWW '19: COMPANION PROCEEDINGS OF THE 2019 WORLD WIDE WEB CONFERENCE, 539-544 (2019).

[53] Philip M. Napoli, On Automation in Media Industries: Integrating Algorithmic Media Production into Media Industries Scholarship, 1 MEDIA INDUSTRIES J. 33 (2014).

[54] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015).

[55] Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO, L. REV. 1671, 1688 (2020).

replicate, obfuscate and entrench historical discriminatory structures, painting them with a veneer of objectivity and fairness while replicating many of their worst aspects.[56]

4.1. AI Bias and Traditional Legal Notions of Discrimination

An early focus of academics and civil society researchers has been around mapping these impacts, and their consequences, from a legal and social perspective. The traditional legal focus on discrimination related to employment has led to particular attention being devoted to the growing use of algorithms to sort and rank potential job applicants.[57] The use of AI in hiring decisions, even if only for triaging potential candidates, poses a concern not only because it can produce discriminatory results, but because it often does so through a facially non-discriminatory decision-making pattern. For example, one algorithm, which was designed to assess potential candidates based on the performance of existing employees, concluded that the two factors which most strongly correlated to strong performance were whether the candidate had played high school lacrosse, and whether their name was Jared.[58] Although neither categorization is discriminatory on a protected ground per se, these kinds of results are obviously going to be strongly associated with protected variables.

Even where an algorithm is specifically prohibited from decision-making based on protected variables, it may nonetheless cultivate stand-in variables as proxies to achieve the same, discriminatory result.[59] As noted in the second section, it is difficult for an algorithm to possess the discriminatory intent that is often required in order to make a legal challenge stick, making it difficult to develop robust structures for legal accountability. Moreover, the prevalence of these proxy characteristics, which may be closely associated with, but distinct from, protected grounds, has the potential to allow decision-makers with prejudicial values to hide deliberate discrimination behind a mask of deniability.[60]

One of the primary benefits of technologies like AI decision-making is its speed and scalability, performing tasks that in an earlier age would have required an army of dedicated staff and a significant investment of time. For example, the use of AI in hiring also has the potential to metastasize the impact of these biases on a massive scale. While there is no question that

[56] It is relevant to note that bias and discrimination are not interchangeable terms, and that there are forms of bias which may be problematic from a design standpoint, but which may not give rise to a valid discrimination complaint because they do not engage with a protected ground or category of persons.

[57] Goldman Sachs, for example, announced in 2016 that it would rely on algorithmic models to automate not only their hiring decisions, but virtually all employee management decisions: Rob Copeland & Bradley Hope, *The World's Largest Hedge Fund Is Building an Algorithmic Model from Its Employees' Brains*, WALL ST. J. (Dec. 22, 2016), https://www.wsj.com/ articles/the-worlds-largest-hedge-fund-is-building-an-algorithmic-model-of-its-founders-brain-1482423694.

[58] See Dave Gershgorn, *Companies Are on the Hook if Their Hiring Algorithms Are Biased*, QUARTZ (Oct. 22, 2018), https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring- algorithms-are-biased.

[59] Piotr Sapiezynski, Avijit Ghosh, Levi Kaplan, Aaron Rieke & Alan Mislove, *Algorithms that "Don't See Color": Measuring Biases in Lookalike and Special Ad Audiences*, https://mislove.org/publications/Lookalike-AIES.pdf.

[60] Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 692-93 (2016).

discrimination in traditional human decision-making systems could also be structurally enabled and enforced, AI imposes an unprecedented level of uniformity and consistency to these decisions, calibrating them to a single standard. If this standard is biased or discriminatory, it can infect entire industries, causing far more harm than a single racist or sexist hiring manager. While an AI decision-maker may be easier to retrain than a human, this requires that one first be able to isolate and diagnose the problem, which is challenging given the complex and opaque way that AI decisions are made.[61]

In considering appropriate policy and regulatory responses to these challenges, scholars of race and technology, such as Safiya Noble, Ruha Benjamin, and Ifeoma Ajunwa, have been at the leading edge of academic thinking around AI and bias, cautioning that without early intervention, the rollout of AI systems across the public and private sectors poses a grave threat to efforts to combat structural inequality and racism.[62]

One prominent theme has been that the challenges posed by biased AI decision-making need to be understood as more than mere technical glitches, which may be resolved through better code, better auditing, or a more judicious selection of training data. Ifeoma Ajunwa, in particular, has argued that the categorization of these challenges as technical problems is fundamentally misguided because there is always a human behind the curtain.[63] Moreover, she argues biased or discriminatory outcomes, even where directly delivered by an algorithm, should be viewed as a legal problem caused by anachronistic approaches towards regulating discriminatory decision-making, such as an overly deferential attitude towards employer choices.[64]

In other words, the spread of algorithms can not only exacerbate and reflect historical biases, but it can also create new opportunities for historical legal deficiencies to be exploited towards discriminatory ends. Solutions that aim to combat discrimination and bias should therefore not only target the problems with the algorithms and their underlying data, but should also aim to rectify these deficiencies in the surrounding legal or policy structure, such as through granting less deference to employers' decision-making which produces discriminatory outcomes.

In a similar vein, Sandra G. Mayson has argued that the challenge from AI technologies is fundamental to their predictive nature, since any predictive system will, by its nature, project the inequities of the past onto the future.[65] On the use of AI in the criminal justice system, Professor Mayson argues that the problem is not just with the design of these systems, but with the very notion of predictive policing. Her response would therefore be to reconsider how we

[61] Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO, L. REV. 1671, 1679 (2020).

[62] *See, e.g.* SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); Ruha Benjamin, *Assessing risk, automating racism* 366 SCIENCE 421 (2019); Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO, L. REV. 1671 (2020).

[63] Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO, L. REV. 1671, 1708 (2020).

[64] *Ibid.*

[65] Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L. J. 2218 (2019).

assess and respond to risk, since these interventions, and how we criminalize certain behaviors and respond to the emergence of criminal patterns as we have defined them, lies at the core of the discriminatory outputs that the criminal justice system generates.

Other legal scholarship has focused more specifically on technical fixes, or at the very least on reconsidering our approach to how AI is developed, implemented, and audited. Anupam Chander, writing in 2017, argued for the institution of a form of "algorithmic affirmative action", which would force a consideration of the disparate impacts of data and design related to categories where discrimination is legally prohibited (i.e., race, age, sex, religion, etc.), and attempt to rectify these impacts through changes to the data or design which return less discriminatory results.[66]

A major challenge with implementing such technical solutions is the lack of public access to accurate information about how these systems were trained and are functioning. In one particularly well known case, an algorithmic recommendation tool meant to guide sentencing, known as COMPAS, was found to be returning results that were biased against Black subjects, flagging them as a significantly greater risk to reoffend.[67] This case is particularly noteworthy in that there was an auditing procedure in place, which found that the system was fair because its overall accuracy rate in terms of predicting recidivism was roughly equivalent between racial groups. The audit neglected to consider that where the system failed it did so by placing Black defendants in a riskier category, and white defendants in a less risky category. Legal scholars have suggested a range of tools aimed and mitigating this specific challenge, including developing and enforcing codes of conduct for the design of AI systems, and enhancing whistleblower protection rules to ensure that internal knowledge about discriminatory systems makes its way into the public realm.[68]

[66] Anupam Chander, *The Racist Algorithm?,* 115 MICH. L. REV. 1023 (2017).

[67] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[68] Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54 (2019).

5. Concluding Recommendations

Challenges related to discrimination and structural bias are by no means a recent phenomenon, and certainly are not unique to AI. The salience of these technologies to the human condition and its ability both to foreground new challenges and to problematize existing and emerging social and legal challenges, mandates serious consideration of how judicial structures should approach challenges that will inevitably flow from the implementation of discriminatory algorithms across the public and private sectors.[69] **Judges should expect that problems related to bias are likely to manifest, in one form or another, across virtually every field where AI decision-making has become popularized.** While every case is unique, and some complaints will bear more merit than others within the context of the prevailing legal framework, judges should keep an open mind towards thinking through which approaches to the law may need to be adapted or reconsidered in light of the transformative impact of these technologies on human decision-making.

[69] Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. Rev. 1 (2004).