Journal of Business & Technology Law

Volume 19 | Issue 2 Article 5

The Sins of the Father: Excising Malignant Bias From Artificial Intelligence

Simon R. Graf

Follow this and additional works at: https://digitalcommons.law.umaryland.edu/jbtl

Recommended Citation

Simon R. Graf, *The Sins of the Father: Excising Malignant Bias From Artificial Intelligence*, 19 J. Bus. & Tech. L. (2023)

Available at: https://digitalcommons.law.umaryland.edu/jbtl/vol19/iss2/5

This Notes & Comments is brought to you for free and open access by the Academic Journals at DigitalCommons@UM Carey Law. It has been accepted for inclusion in Journal of Business & Technology Law by an authorized editor of DigitalCommons@UM Carey Law. For more information, please contact smccarty@law.umaryland.edu.

THE SINS OF THE FATHER: EXCISING MALIGNANT BIAS FROM ARTIFICIAL INTELLIGENCE

SIMON R. GRAF*

> *I am worried that algorithms are getting too prominent in the world. It started out that computer scientists were worried nobody was listening to us. Now I'm worried that too many people are listening.*[1]

INTRODUCTION

Artificial Intelligence (AI) has permeated nearly every pore of our society,[2] from autonomous vehicles[3] to digital assistants[4] to facial recognition

© 2024 Simon R. Graf.

* J.D. Candidate 2025, University of Maryland Francis King Carey School of Law. I would like to thank Professor William J. Moon for his guidance and counsel; Tom, Penny, and Marie Graf, Rob Coman, and Reema Sood, for their endless love, support, and encouragement; Matt and Priya Cherry, Jason Barbour, Lisa Auld, Wayne Elseth, and Helen Cramer, for taking me under their respective wings; the Honorable Douglas R. M. Nazarian, Theresa Breslin, Sangeetha Kannan, and Victoria Szczawinski Trocchia, for their kind and thoughtful instruction; and the editors of the Journal of Business & Technology Law for their feedback, suggestions, and attention to detail.

[1] Siobhan Roberts, *The Yoda of Silicon Valley*, N.Y. TIMES (Dec. 17, 2018), https://www.nytimes.com/2018/12/17/science/donald-knuth-computers-algorithms-programming.html; RUHA BENJAMIN, RACE AFTER TECHNOLOGY: ABOLITIONIST TOOLS FOR THE NEW JIM CODE 16 (2019) (contextualizing Donald Knuth's thoughts on AI); *see infra* Conclusion.

[2] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. OF MACH. LEARNING RSCH. 77 (2018), https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[3] Troy Griggs & Daisuke Wakabayashi, *How a Self-Driving Uber Killed a Pedestrian in Arizona*, N.Y. TIMES (Mar. 21, 2018), https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html; Aarian Marshall & Alex Davies, *Uber's Self-Driving Car Didn't Know Pedestrians Could Jaywalk*, WIRED (Nov. 5, 2019, 9:22 PM), https://www.wired.com/story/ubers-self-driving-car-didnt-know-pedestrians-could-jaywalk.

[4] Mark West, Rebecca Kraut & Chew Han Ei, *The Rise of Gendered AI and Its Troubling Repercussions*, UNESCO & EQUALS SKILLS COALITION 90 (2019), https://doi.org/10.54675/RAPC9356.

systems.[5] AI is a highly technical discipline,[6] the inner workings of which are often opaque,[7] withheld from the public on a proprietary basis,[8] or otherwise inaccessible.[9] Academically speaking, "Artificial Intelligence" is the study of how to make computers emulate actions and behaviors that we associate with human thinking, such as "decision-making, problem solving, learning,"[10] "us[ing] language, form[ing] abstractions and concepts, solv[ing the] kinds of problems now reserved for humans, and improv[ing] themselves."[11] Practically speaking, "AI" is an umbrella term encompassing many distinct but related models for automating tasks and decisions that would otherwise be assigned to humans. Scientists study AI for many different reasons, including to gain a greater philosophical understanding of human thought;[12] as a purely academic exploration of computer capabilities;[13] to simplify or automate complex, rote, repetitive, or otherwise unpalatable tasks or decisions;[14] and to develop systems that remove human subjectivity from decision-making.[15]

A system that makes a decision or judgment based, at least in part, on the output of an AI algorithm is often referred to as an Automated Decision System[16] (ADS).[17] Although some varieties of AI are characterized by

[5] Elizabeth McClellan, Comment, *Facial Recognition Technology: Balancing the Benefits and Concerns*, 15 J. BUS. & TECH. L. 363, 363-64 (2020); Khari Johnson, *How Wrongful Arrests Based on AI Derailed 3 Men's Lives*, WIRED (Mar. 7, 2022, 7:00 AM), https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives; Kashmir Hill, *Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match*, N.Y. TIMES (Jan. 6, 2021), https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html.

[6] The field of AI was founded upon the "conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 17 (Stuart J. Russell & Peter Norvig eds., 3d ed. 2010).

[7] BENJAMIN, *supra* note 1, at 15 (discussing Silicon Valley's "ruthless code of secrecy").

[8] Sarah Myers West, Meredith Whittaker & Kate Crawford, *Discriminating Systems: Gender, Race and Power in AI*, AI NOW INST. 19 (2019), https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2.

[9] "Opaque and invisible models are the rule, and clear ones very much the exception." CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 28 (2016).

[10] RUSSELL & NORVIG, *supra* note 6, at 2.

[11] *Id.* at 17.

[12] *Id.* at 4-9, 10-13.

[13] *Id.* at 2-4.

[14] *Id.* at 9-10, 13-14.

[15] *Id.* at 15, 17-18.

[16] Some sources instead refer to this concept as an "Algorithmic Decision System," but these terms are interchangeable.

[17] An alternative definition describes an automated decision system as:

their ability to "learn," an algorithm need not be capable of learning to fall into the category of an ADS. Indeed, the U.S. government has defined the term "Automated Decision System" to mean "any system, software, or process (including one derived from Machine Learning, **statistics**, or **other data processing** or artificial intelligence techniques and excluding passive computing infrastructure) **that uses computation**, the result of which serves as a basis for a decision or judgment."[18] An ADS, then, can be as simple as one or more computations used to make some determination.

A thoughtful implementation of AI has great potential to simplify and expedite routine tasks and produce more consistent results compared to humans.[19] Indeed, ADS are already employed to guide professionals in

> any software, system, or process that aims to automate, aid, or replace human decision-making. Automated decision systems can include both tools that analyze datasets to generate scores, predictions, classifications, or some recommended action(s) that are used by agencies to make decisions that impact human welfare, and the set of processes involved in implementing those tools.

Rashida Richardson, ed., *Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force* 20, AI NOW INST. (Dec. 4, 2019), https://ainowinstitute.org/publication/confronting-black-boxes-a-shadow-report-of-the-new-york-city-automated [hereinafter *Confronting Black Boxes*]. The definition further explains that impacting human welfare "includes but is not limited to decisions that affect sensitive aspects of life such as educational opportunities, health outcomes, work performance, job opportunities, mobility, interests, behavior, and personal autonomy." *Id.* at n.37.

[18] Algorithmic Accountability Act of 2022, S. 3572, 117th Cong. § 2(2) (2022) (emphasis added); Rashida Richardson, Jason M. Schultz & Vincent M. Southerland, *Litigating Algorithms 2019 US Report: New Challenges to Government Use of Algorithmic Decision Systems*, AI NOW INST. 7 (Sept. 2019), https://ainowinstitute.org/publication/litigating-algorithms-2019-u-s-report-2.

[19] NAT'L INST. OF STANDARDS AND TECH., NIST SPECIAL PUB. 1270, TOWARDS A STANDARD FOR IDENTIFYING AND MANAGING BIAS IN ARTIFICIAL INTELLIGENCE 20 (Mar. 2022), https://doi.org/10.6028/NIST.SP.1270 [hereinafter MANAGING BIAS IN AI].

healthcare,[20] criminal justice,[21] actuarial science,[22] education,[23] employment,[24] and more. By contrast, numerous studies,[25] lawsuits,[26] and high-

[20] *See* Tom Simonite, *A Health Care Algorithm Offered Less Care to Black Patients*, WIRED (Oct. 24, 2019, 2:00 PM), https://www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care (AI in healthcare favored White patients); Heidi Ledford, *Millions of Black People Affected by Racial Bias in Health-care Algorithms*, 574 NATURE 608, 608-09 (Oct. 24, 2019), https://www.nature.com/articles/d41586-019-03228-6 ("An algorithm widely used in US hospitals to allocate health care to patients has been systematically discriminating against black people."); Darshali A. Vyas et al., *Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms*, 383 NEW ENG. J. OF MED. 874 (Aug. 27, 2020), https://doi.org/10.1056/NEJMms2004740 ("Many . . . race-adjusted algorithms guide decisions in ways that may direct more attention or resources to white patients than to members of racial and ethnic minorities."); Laleh Seyyed-Kalantari et al., *Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-served Patient Populations*, 27 NATURE MED. 2176 (Dec. 2021), https://doi.org/10.1038/s41591-021-01595-0 (tension between pros and cons of AI is "particularly pressing in healthcare"); Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCI. 447 (Oct. 25, 2019), https://doi.org/10.1126/science.aax2342 ("The U.S. health care system uses commercial algorithms to guide health decisions.").

[21] *See* Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (exposing flaws with risk assessment algorithms in criminal justice); Carrie Johnson, *Justice Department Works to Curb Racial Bias in Deciding Who's Released from Prison*, NPR (Apr. 19, 2022, 12:28 PM), https://www.npr.org/2022/04/19/1093538706/justice-department-works-to-curb-racial-bias-in-deciding-whos-released-from-pris [hereinafter *Justice Department Works to Curb Racial Bias*] (reporting recidivism risk assessment tool bias); Carrie Johnson, *Flaws Plague a Tool Meant to Help Low-risk Federal Prisoners Win Early Release*, NPR (Jan. 26, 2022, 5:00 AM), https://www.npr.org/2022/01/26/1075509175/justice-department-algorithm-first-step-act [hereinafter *Flaws in the First Step Act*] ("[T]he [DOJ] said its algorithmic tool for assessing the risk that a person in prison would return to crime produced uneven results.").

[22] *See* Jay Vadiveloo, *Model Behavior: Applications of Artificial Intelligence in Actuarial Science,* CONTINGENCIES, Nov.-Dec. 2019, at 21, https://view.publitas.com/ba55d288-8598-4c1a-8a1f-b0e6140f5b5a/cont_2019_1112/page/1 (discussing potential benefits of incorporating AI into actuarial science).

[23] *See* Todd Feathers, *Major Universities Are Using Race as a "High Impact Predictor" of Student Success*, THE MARKUP (Mar. 2, 2021, 8:00 AM), https://themarkup.org/machine-learning/2021/03/02/major-universities-are-using-race-as-a-high-impact-predictor-of-student-success (describing problems with algorithms used by universities to predict the risk that a student will drop out); Lydia X. Z. Brown, *How Automated Test Proctoring Software Discriminates Against Disabled Students*, CTR. FOR DEMOCRACY & TECH. (Nov. 16, 2020), https://cdt.org/insights/how-automated-test-proctoring-software-discriminates-against-disabled-students (explaining how automated test proctoring software discriminates against disabled students); NAT'L DISABLED L. STUDENTS ASS'N, REPORT ON CONCERNS REGARDING ONLINE ADMINISTRATION OF BAR EXAMS 20-21 (July 2020), https://ndlsa.org/wp-content/uploads/2020/08/NDLSA_Online-Exam-Concerns-Report1.pdf (more software discrimination against disabled students).

[24] *See* Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 10, 2018, 7:04 PM), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-

publicity gaffes[27] have illustrated that AI is only as "intelligent" as its developers train it to be. There is a common perception among the public—often invoked as an alluring justification for delegating public interest decisions to ADS technologies—that because AI is consistent, objective, and data-driven, it is inherently fair and bias-free.[28] The unfortunate reality is that AI is capable of being objective and biased at the same time, and when deployed prematurely in high-stakes settings, these systems can "perpetuate harms more quickly, extensively, and systematically than human and societal biases on their own."[29] To make matters worse, there is no way to guarantee that an algorithm is not biased—or will not

idUSKCN1MK08G (employment search algorithm discriminated against women); U.S. Dept. of Just. Civ. Rts. Div., *Algorithms, Artificial Intelligence, and Disability Discrimination in Hiring* 1-2 (May 12, 2022), https://www.ada.gov/resources/ai-guidance (explaining how AI can lead to discrimination against disabled applicants in hiring); Miranda Bogen, *All the Ways Hiring Algorithms Can Introduce Bias*, Harv. Bus. Rev. (May 6, 2019), https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias ("Unfortunately, we found that most hiring algorithms will drift toward bias by default.").

[25] Andrew Blair-Stanek et al., *GPT-4's Law School Grades: Con Law C, Crim C-, Law & Econ C, Partnership Tax B, Property B-, Tax B* 1-2 (May 9, 2023), http://dx.doi.org/10.2139/ssrn.4443471 (subjecting GPT-4 to a battery of law school exams and studying the results); AI Now Inst., *The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term* 6-10 (Sept. 22, 2016), https://artificialintelligencenow.com/media/documents/AINowSummaryReport_3_RpmwKHu.pdf (exploring how AI systems might "contribute to unfair bias and discrimination"); *Confronting Black Boxes*, *supra* note 17, at 7-8 (studying the successes and failures of a New York City pilot project exploring government uses of AI).

[26] Richardson, Schultz & Southerland, *supra* note 18, at 5-11, 13-15, 19-26, 28-32 (examining numerous lawsuits across the country and internationally involving AI).

[27] Paresh Dave, *AI Algorithms Are Biased Against Skin With Yellow Hues*, WIRED (Oct. 3, 2023, 7:00 AM), https://www.wired.com/story/ai-algorithms-are-biased-against-skin-with-yellow-hues; Tom Simonite, *When It Comes to Gorillas, Google Photos Remains Blind*, WIRED (Jan. 11, 2018, 7:00 AM), https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind [hereinafter *When It Comes to Gorillas*]; James Vincent, *Google 'Fixed' its Racist Algorithm by Removing Gorillas From its Image-labeling Tech*, The Verge (Jan. 12, 2018, 10:35 AM), https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai; Andrew Thompson, *Google's Sentiment Analyzer Thinks Being Gay Is Bad*, Vice (Oct. 25, 2017, 1:00 PM), https://www.vice.com/en/article/j5jmj8/google-artificial-intelligence-bias; Nat'l Inst. of Just., 2021 Review and Revalidation of the First Step Act Risk Assessment Tool 3-4 (Dec. 2021), https://nij.ojp.gov/library/publications/2021-review-and-revalidation-first-step-act-risk-assessment-tool; Nat'l Inst. of Standards and Tech., NIST Interagency or Internal Rep. 8280, Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects 1-4, 6, 14 (Dec. 2019), https://doi.org/10.6028/NIST.IR.8280.

[28] Managing Bias in AI, *supra* note 19, at 33; Exec. Off. of the President, Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights 6 (May 2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[29] Managing Bias in AI, *supra* note 19, at 33.

become biased in the future.[30] To build a safe, equitable, and ethical foundation for public-facing AI algorithms, we must regulate the four interdependent "cornerstones" of trustworthy ADS: fairness,[31] transparency,[32] accountability,[33] and sustainability.[34]

This paper will explore the causes and discriminatory effects of algorithmic bias[35] in AI and will propose a regulatory model to reduce and remedy the propagation of biased AI.[36] First, Section I will examine the origins of the three types of algorithmic bias, as identified by the National Institute of Standards and Technology (NIST). Next, Section II will detail the two distinct manifestations of algorithmic bias and their respective consequences. Finally, Section III will propose a regulatory framework for how to protect vulnerable populations from algorithmic bias, mitigate adverse effects, and provide legal recourse for those affected.

[30] *Id.* at ii.

[31] *See* Gregory S. Nelson, *Bias in Artificial Intelligence*, 80 N.C. MED. J. 220, 221 (July 1, 2019), https://doi.org/10.18043/ncm.80.4.220 (emphasizing the importance of fairness in AI algorithms).

[32] *See id.* at 221 (emphasizing the importance of transparency in the development of AI algorithms); *see also* Lucy Vasserman & John Cassidy, *Increasing Transparency in Perspective's Machine Learning Models*, JIGSAW (Jan. 30, 2019), https://medium.com/jigsaw/increasing-transparency-in-machine-learning-models-311ee08ca58a (emphasizing importance of transparency in operationalized AI models).

[33] *See* Nelson, *supra* note 31, at 221 (emphasizing the importance of accountability in the development of AI algorithms).

[34] *See* WHITE HOUSE OFF. OF SCI. AND TECH. POL'Y, *The Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People* 50 (Oct. 2022), https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf [hereinafter *The Blueprint for an AI Bill of Rights*] (emphasizing the importance of ongoing assessment and maintenance in combatting automation bias).

[35] "Algorithmic bias" is "a term used to describe systematic and repeatable errors in a computer system" that often result in discriminatory treatment of legally protected groups and traits, such as race and gender. *See* Maya C. Jackson, Comment, *Artificial Intelligence & Algorithmic Bias: The Issues with Technology Reflecting History & Humans*, 16 J. BUS. & TECH. L. 299, 300 (2021).

[36] It is important to note that some types of algorithmic bias in AI systems are incorporated deliberately but intended to be *beneficial* to users. Namely, algorithms designed to "creat[e] positive experiences for online shopping or identifying content of interest" in the context of advertising and recommendation engines. *See* MANAGING BIAS IN AI, *supra* note 19, at 3.

I. THE PERNICIOUS PREDISPOSITION OF AI: DE FACTO DISCRIMINATION

AI has the potential to revolutionize our society, and indeed, it already has.[37] But AI has a fatal flaw: if not developed with the utmost care and attention to detail, it is dangerously predisposed to imparted bias.[38] Even when such care is shown, there is no way to guarantee that an algorithm will be completely bias-free.[39] This bias can be imparted either intentionally or unintentionally—and can even develop over time as AI is used in practice—but in most cases it is an inadvertent side effect.[40] As NIST explains, algorithmic bias is often an unavoidable byproduct of the development process:

> *The teams involved in AI system design and development bring their cognitive biases, both individual and group, into the process. Bias is prevalent in the assumptions about which data should be used, what AI models should be developed, where the AI system should be placed — or if AI is required at all. . . . Biases impacting human decision making are usually implicit and unconscious, and therefore unable to be easily controlled or mitigated.*[41]

A lack of diversity in the tech industry—and in AI, specifically—translates into development teams that share a demographic composition: mostly White; mostly male.[42] Teams of developers with a shared demographic makeup have a higher likelihood of baking the same flavors of bias into their AI.[43] Although algorithmic bias can adversely affect any

[37] *See* Nithesh Naik et al., *Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?*, 9 FRONTIERS IN SURGERY 1, 2 (Mar. 14, 2022), https://doi.org/10.3389/fsurg.2022.862322 ("If harnessed effectively, such AI-clinician cooperation . . . . can provide healthcare offerings in diagnosis, drug discovery, epidemiology, personalized care, and operational efficiency.").

[38] *See* Nelson, *supra* note 31, at 220 ("Bias is a reflection of the data [that] algorithm authors choose to use, as well as their data blending methods, model construction practices, and how results are applied and interpreted. That is to say, these processes are driven by human judgments.").

[39] MANAGING BIAS IN AI, *supra* note 19, at ii.

[40] *Id.* at 3.

[41] *Id.* at 5.

[42] West, Whittaker & Crawford, *supra* note 8, at 3, 5, 10-11.

[43] *See* Bo Cowgill et al., *Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics*, NAVIGATING THE BROADER IMPACTS OF AI RSCH. WORKSHOP AT THE 34TH CONF. ON NEURAL INFO. PROCESSING SYS. 4 (Dec. 4, 2020),

population, the one-dimensional state of the industry preordains AI with biases that, curiously, tend not to harm White males. Rather, biased AI most commonly discriminates based on race, gender, age, socioeconomic stratum, disability, religious affiliation, and national origin, as well as other protected groups.[44] In other words, our most vulnerable and marginalized populations bear the consequences of algorithmic bias imparted by predominantly White male developers. To add insult to injury, algorithmic bias is often self-reinforcing, which amplifies the fallout and traps victims in a perpetual cycle of AI-administered oppression.[45] And when the output of one ADS is passed to another, a single adverse determination can have a cascading effect, diffusing into nearly every layer of a victim's life.[46] These instances of algorithmic bias can be divided into three categories: systemic bias, statistical bias, and human bias.[47]

A. Types of Bias

Systemic bias originates at the institutional level, where there exist procedures and practices that result in "certain social groups being advantaged or favored and others being disadvantaged or devalued."[48] Common examples of systemic bias include institutionalized racism and sexism. A historic example of systemic bias is Kodak's "Shirley Cards," produced from the 1950s until the 1990s, were aids used by photo labs to calibrate

https://doi.org/10.48550/arXiv.2012.02394 ("However, we do find that prediction errors are correlated within demographic groups, particularly gender. Specifically, two male programmers' prediction errors are more likely to be correlated with each other. A team or ensemble approach that averages across two male programmers will effectively double down these errors.").

[44] *The Blueprint for an AI Bill of Rights*, *supra* note 34, at 5.

[45] *See* Danielle Ensign et al., *Runaway Feedback Loops in Predictive Policing*, 81 PROC. MACH. LEARNING RSCH. 1, 1-3 (2018), https://proceedings.mlr.press/v81/ensign18a/ensign18a.pdf (describing how a feedback loop in predictive policing tools causes police to erroneously uproot the same neighborhoods).

[46] Cathy O'Neil describes how this cascading effect can manifest:

> Poor people are more likely to have bad credit and live in high-crime neighborhoods, surrounded by other poor people. Once the dark universe of [ADS] digests that data, it showers them with predatory ads for subprime loans or for-profit schools. It sends more police to arrest them, and when they're convicted it sentences them to longer terms. This data feeds into other [ADS], which score the same people as high risks or easy targets and proceed to block them from jobs, while jacking up their rates for mortgages, car loans, and every kind of insurance imaginable. This drives their credit rating down further, creating nothing less than a death spiral of modeling.

O'NEIL, *supra* note 9, at 199-200.

[47] MANAGING BIAS IN AI, *supra* note 19, at 6, 9.

[48] *Id.* at 6.

film exposure.[49] The cards presented a photo of a White woman surrounded by rectangles of various colors, and were intended to help photo lab technicians calibrate the color of a photo to achieve a "normal" appearance before printing.[50] "Since the model's skin was set as the norm," Benjamin writes, "darker skinned people in photographs would be routinely underexposed."[51]

Statistical bias is caused when a sample set of data does not accurately represent the broader population.[52] NIST explains that "these biases . . . often arise when algorithms are trained on one type of data and cannot extrapolate beyond those data."[53] For example, in 2015, the Intelligence Advanced Research Projects Activity (IARPA) released a 500-subject dataset called "IARPA Janus Benchmark A" (or IJB-A), which was intended to be "the most geographically diverse set of collected faces" at that time.[54] "Preliminary analysis of the IJB-A . . . benchmarks," Buolamwini and Gebru write, "revealed overrepresentation of lighter males, underrepresentation of darker females, and underrepresentation of darker individuals in general."[55] Using the Fitzpatrick Skin Type classification system as a guide, Buolamwini and Gebru found that 79.6% of the IJB-A dataset featured lighter skin tones, while only 20.4% featured darker skin tones.[56] Buolamwini and Gebru tested three commercial AI products built to identify the gender of a photographed person. Confirming a finding of statistical bias, they noted that "all algorithms perform[ed] worse on female and darker subjects when compared to their counterpart male and lighter subjects."[57]

Human bias—sometimes referred to as *implicit* bias—"reflect[s] systematic errors in human thought based on a limited number of heuristic[58] principles" and "tend[s] to relate to how an individual or group perceives information . . . to make a decision or fill in missing or unknown

[49] BENJAMIN, *supra* note 1, at 103.

[50] Mandalit del Barco, *How Kodak's Shirley Cards Set Photography's Skin-Tone Standard*, NPR (Nov. 13, 2014, 3:45 AM), https://www.npr.org/2014/11/13/363517842/for-decades-kodak-s-shirley-cards-set-photography-s-skin-tone-standard.

[51] BENJAMIN, *supra* note 1, at 103-04.

[52] MANAGING BIAS IN AI, *supra* note 19, at 9.

[53] *Id.*

[54] Buolamwini & Gebru, *supra* note 2, at 3.

[55] *Id.* at 5.

[56] *Id.* at 7.

[57] *Id.* at 10.

[58] A heuristic is an "adaptive mental shortcut[]" that can help reduce the complexity of judgments and choices. However, because heuristics "cut corners," so to speak, they can also contribute to cognitive bias. *See* MANAGING BIAS IN AI, *supra* note 19, at 9.

information."[59] One highly publicized example is the "antidiversity" manifesto written in 2017 by former Google employee, James Damore. In it, Damore effectively argued—among other things—that "women are psychologically inferior and incapable of being as good at software engineering as men."[60] Although human bias can take the form of individualized prejudice, in a broader sense it would be more precise to think of it as the outcome of applied logical fallacies and other cognitive and perceptual biases.[61]

B. The Origins of Algorithmic Bias

The study of AI is composed of many distinct (and often overlapping) subfields, such as deep learning, neural networks, natural language processing, and more.[62] One of the most prevalent and best-known subfields of AI is called Machine Learning (ML). Machine Learning is a type of AI characterized by its "ability to automatically learn and improve on the basis of data or experience, without being explicitly programmed."[63] In other words, Machine Learning algorithms are not given instructions on how to make decisions. Rather, they are little more than pattern recognition algorithms with the added ability to classify (*i.e.*, categorize) new information based on patterns they have identified.[64] Although bias can manifest in any type of AI, the following sections are in the context of Machine Learning.

1. Black Box

A Machine Learning algorithm's ability to classify new information is developed through a "training phase," during which developers provide as input to the algorithm a pre-selected and pre-categorized dataset.[65] For example, the training data for an algorithm intended to recognize animals in pictures might consist of hundreds, thousands, or even millions of

[59] *Id.*

[60] SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM 2 (2018).

[61] Including, for example, confirmation bias, availability bias, and anchoring bias. *See* MANAGING BIAS IN AI, *supra* note 19, at 9.

[62] These subfields will not be discussed in this paper.

[63] National Artificial Intelligence Initiative Act of 2020, 15 U.S.C. § 9401(11) (2020).

[64] *See* CHRISTOPHER M. BISHOP, PATTERN RECOGNITION AND MACHINE LEARNING 1 (2006) ("The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories.").

[65] *Id.* at 2.

images of animals—each pre-labeled with the type (or types) of animal in the picture. Feeding these training data into the algorithm allows it to draw inferences about what an image of a particular animal "looks like."[66] In practice, this algorithm would be given unlabeled images of animals and asked to reverse the recognition process. That is, given an unlabeled image as input, the algorithm would compare the characteristics of that image to its "understanding" of what each animal "looks like" and would return the most likely animal (or animals) as output.

Machine Learning algorithms can be incredibly powerful,[67] but the road to utility is riddled with pitfalls. First, problems with an algorithm can be hard to spot and harder still to fix. Because ML algorithms produce their own internal criteria for classifying data, examining their inner workings to diagnose a problem is not as simple as inspecting the source code.[68] As noted earlier, ML algorithms are not programmed with discrete evaluation instructions. Rather, algorithms develop their own classification criteria through their interactions with a set of training data. Generally, these developed criteria are not "examinable."[69] That is, the algorithm has no means of communicating its classification criteria to developers in plain language, source code, or otherwise. Thus, even when

[66] *See* Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 6 (2019), https://scholarship.law.upenn.edu/faculty_scholarship/2123 ("Machine Learning's value derives from its ability to learn for itself how to detect useful patterns in massive data sets and put together information in ways that yield remarkably accurate predictions or estimations.").

[67] And with great power comes great responsibility. *See* Stan Lee, Steve Ditko, & Stan Goldberg, *Spider-Man!*, AMAZING FANTASY, VOLUME ONE 15, at 11 (Atlas Magazines, Inc. Aug. 1962) *available at* https://archive.org/details/Amazing_Fantasy_vol1_15_201607/page/n1/mode/2up (last visited Feb. 11, 2024); *see generally* Britton Payne, *Comic Book Citation Format*, 16 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 1017, 1017-20 (2006) (proposing a legal citation format for comic books previously unaddressed by THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia L. Rev. Ass'n et al. eds., 21st ed. 2020)).

[68] Coglianese and Lehr explain the difficulty of identifying problems with ML classification criteria:

> Even if analysts could discover the inter-variable relationships that a machine-learning algorithm keys in on, they cannot overlay any causal inferences onto those relationships. In other words, they cannot say that a relationship between an input variable and the output variable is causal in nature. In fact, some of the patterns that are predictively useful might not be causal at all, and some may be so non-intuitive as to have never occurred to humans—perhaps, say, if the third letter of a tax filer's last name helps in forecasting cases of tax fraud.

Coglianese & Lehr, *supra* note 66, at 17.

[69] *See id.* at 5 n.8 (citing JUDEA PEARL & DANA MACKENZIE, THE BOOK OF WHY: THE NEW SCIENCE OF CAUSE AND EFFECT 359 (2018)) (explaining that with ML techniques, "the programmer has no idea what computations [the algorithm] is performing or why they work").

developers determine that an algorithm is not functioning as intended, fixing the problem is rarely as simple as changing a few lines of code.[70]

2. Spurious Suppositions

Next, just as human thought is susceptible to spurious correlations, so too are ML algorithms.[71] These types of issues can be incredibly difficult to track down without exhaustive testing because (a) developers cannot inspect *why* an algorithm makes certain determinations, and (b) developers are unlikely to question the efficacy of such determinations if the classification accuracy is high.[72] To use the animal classifier example, an algorithm might erroneously conclude that any image with a sky-blue background is a bird. An inference of this type could classify images with high accuracy, even though it is based on an unintended criterion.[73] Developers often fail to consider the presence of a spurious correlation due to confirmation bias—in other words, the algorithm appears to be working as intended because the output matched the developers' expectation.[74]

In a real-world example, a German public radio station evaluated a system created to assist companies with the hiring process that created a personality profile based on "tone of voice, language, gestures, and facial expressions."[75] Their testing showed that "the AI system was easily manipulated by superficial changes to its inputs, awarding candidates higher scores when they wore glasses or when a bookshelf was in the background, diminishing claims that the system analyzed human expressions, and raising concerns about shortcut learning."[76]

[70] *See* Cliff Kuang, *Can A.I. Be Taught to Explain Itself?*, N.Y. TIMES MAG. (Nov. 21, 2017), https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html (explaining "that artificial intelligences often excel by developing whole new ways of seeing, or even thinking, that are inscrutable to us").

[71] MANAGING BIAS IN AI, *supra* note 19, at 24.

[72] *See id.* at 25 ("Measuring whether the patterns identified by these applications are real or a result of spurious correlations is difficult.").

[73] *Id.* at 19.

[74] *See id.* at 50 (defining confirmation bias as "a cognitive bias where people tend to prefer information that aligns with, or confirms, their existing beliefs"); *id.* at 27 ("Even among experts, data-driven technologies can exacerbate confirmation bias, particularly when they are implicitly guided by expected outcomes.").

[75] *Id.* at 24.

[76] *Id.*

3. Deviant Data

Another common source of bias is the dataset used to train an algorithm, and many issues with such a dataset can be traced back to human judgment.[77] Before an algorithm can be trained, a dataset must first be procured. Development teams can either (a) create a new dataset tailormade to fit the purpose of the algorithm they are developing, or (b) use an existing dataset that was likely created for a different purpose. Both options are susceptible to bias.

When creating a new dataset, developers' own implicit biases influence the "data selection, curation, preparation and analysis processes."[78] The developers' decisions during data selection, curation, and preparation "affect who and what gets counted, and who and what does not get counted," and developers may decide what to include and exclude in a way that is consistent with their own views and beliefs.[79] "Our own values and desires influence our choices, from the data we choose to collect to the questions we ask. Models are opinions embedded in mathematics."[80]

It is easier, faster, and cheaper to use an existing dataset when available. But choosing a dataset simply because it is accessible introduces the possibility that the data do not accurately or proportionally represent[81] the target population.[82] As a result, the algorithm's "understanding" of classification criteria becomes skewed. This problem can cascade when unrepresentative datasets are reused for some other purpose. This removes the dataset from the social and temporal context within which it was assembled.[83] According to NIST, "[d]isadvantaged groups including indigenous populations, women, and disabled people are consistently underrepresented," and when a dataset has been removed from its original context and repurposed, it can contribute to discrimination against underrepresented groups.[84]

Moreover, research has found that developers with shared demographic characteristics (such as gender or race) are more likely to impart

[77] Cowgill et al., *supra* note 43, at 3.

[78] MANAGING BIAS IN AI, *supra* note 19, at 19.

[79] *Id.*

[80] O'NEIL, *supra* note 9, at 21.

[81] To add insult to injury, NIST cautions that "even if datasets are representative[,] they may still exhibit biases or improperly utilize protected attributes, which in turn may lead to discrimination." Validating and sustaining dataset fairness is a fraught but essential undertaking. *See* MANAGING BIAS IN AI, *supra* note 19, at 30.

[82] *Id.* at 15.

[83] *Id.* at 16.

[84] *Id.*

the same implicit biases in their work.[85] In light of the demographic distribution among AI developers, this is a problematic correlation:

> *Recent studies found only 18% of authors at leading AI conferences are women, and more than 80% of AI professors are men. This disparity is extreme in the AI industry: women comprise only 15% of AI research staff at Facebook and 10% at Google. . . . [O]nly 2.5% of Google's workforce is black, while Facebook and Microsoft are each at 4%.*[86]

Additionally, an analysis conducted in 2017 found that the percentage of technical employees who are women was only 23% at Apple, 20% at Google, and 17.5% at Microsoft.[87] And another report estimated that fewer than 1% of job applications for expert AI and data science positions were submitted by women.[88]

"[L]arge scale AI systems are developed almost exclusively in a handful of technology companies and a small set of elite university laboratories, spaces that in the West tend to be extremely white, affluent, technically oriented, and male.[89] These are also spaces that have a history of problems [with] discrimination, exclusion, and sexual harassment."[90] It is unsurprising, then, that algorithmic bias "mirrors and replicates existing structures of inequality in society."[91] Likewise, those who *benefit* from such bias are generally "those already in positions of power, who again tend to be white, educated, and male."[92]

It is no coincidence that such an uncomfortably homogenous industry has been failing to attract a more diverse pool of applicants. Data from a 2010-2012 survey conducted by the American Institute for Economic Research found that female software developers on average earned less than White women and White, Black, and Asian men; and Latina software developers earned up to 20% less than White male developers.[93]

[85] Cowgill et al., *supra* note 43, at 4.

[86] West, Whittaker & Crawford, *supra* note 8, at 3.

[87] West, Kraut & Chew, *supra* note 4, at 102.

[88] *Id.* at 102-03.

[89] *See* West, Whittaker & Crawford, *supra* note 8, at 11 ("Machine vision researcher and co-founder of Black in AI, Timnit Gebru, said that when she first attended the preeminent machine learning conference NeurIPS in 2016, she was one of six black people – out of 8,500 attendees.").

[90] *Id.* at 6.

[91] *Id.*

[92] *Id.* at 7.

[93] *Id.* at 13.

If a lack of industry diversity results in teams of predominantly White male developers embedding implicit bias in algorithms, then diversity, equity, and inclusion (commonly known as "DEI") initiatives might just be the stone to kill these two birds.[94] NIST supports this notion, suggesting that "ensuring that individuals involved in training, testing, and deploying the system have a diversity of experience, expertise and backgrounds is a critical risk mitigant that can help organizations manage the potential harms of AI."[95] A development team composed of diverse individuals is better equipped to understand how the algorithm or ADS is likely to affect users with a variety of different backgrounds, how those users might engage with the system, how the system could harm or benefit different users and groups, and how the system might affect populations outside the immediate group of intended users.[96]

Forming a diverse team of developers is an important step in the right direction, but developers are not the only voices that matter. The development team—and the organization as a whole—should consult experts and stakeholders within the target population and across social divides that could be affected by bias—such as race, gender, age, and persons with disabilities.[97] After all, "[t]echnology or datasets that seem non-problematic to one group may be deemed disastrous by others."[98] And, at least for those AI systems that iterate and evolve, stakeholder consultations should occur with reasonable regularity to ensure that changing system conditions have not materially altered the system's behavior.[99]

Finally, datasets used for ML applications are, in essence, attempts to reduce objects, abstract concepts, human subjectivity, and ambiguity to an objectively measurable form.[100] This essential step is part of what enables ML algorithms to recognize patterns and classify new data. However, much is lost in translation. This information loss, called "flattening," is an inevitable part of representing "complex human phenomena with mathematical models" and "comes at the cost of disentangling the context

[94] MANAGING BIAS IN AI, *supra* note 19, at 36-37.

[95] *Id.* at 37.

[96] *Id.*

[97] *See* Lucy Vasserman et al., *Unintended Bias and Identity Terms*, JIGSAW (Mar. 9, 2018), https://medium.com/jigsaw/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23 ("[D]iverse points of view make discussions better."); MANAGING BIAS IN AI, *supra* note 19, at 36.

[98] MANAGING BIAS IN AI, *supra* note 19, at 36.

[99] *Id.*

[100] *Id.* at 12.

necessary for understanding individual and societal impact and contributes to a fallacy of objectivity."[101]

To use the animal classification example, labeling each image with the type of animal pictured may be sufficient for a simple purpose, but a label such as "bird" fails profoundly to represent the concept of a bird. And "[o]nce the context has been removed . . . it is difficult to get it back, leading AI models to learn from inexact representations."[102] Adding labels to increase the depth of the model (*e.g.*, labeling the image with "bird," "wings," and "beak") can help mitigate some of the effects of flattening, but (a) information and context is always lost no matter how deep the model,[103] and (b) the more complex the model, the greater the risk of bias. After all, developers ultimately decide "who and what gets counted, and who and what does not get counted."[104] Compounding this risk is that flattening adds ambiguity to human interpretation of an AI model's output and makes it more difficult to track down the precise causes of bias.[105]

4. Misapplication & Misuse

Deployment bias and emergent bias are two sides of the same coin: both are introduced through misuse. Deployment bias occurs when an AI model is used in *unintended ways*, whereas emergent bias arises when an AI model is used in *unanticipated contexts*.[106] Both biases are the result of "concept drift," which is when the original intent or idea for an AI model changes after deployment "as the application is repurposed or used in unforeseen ways, and in settings or contexts for which it was not originally intended."[107] Concept drift is a common cause of disparities in ADS performance between laboratory settings and deployment, not to mention that when an algorithm is repurposed or used in a new context, new risks are introduced that can cause further performance disparities.[108]

[101] *Id.*

[102] *Id.* at 18.

[103] *See* O'NEIL, *supra* note 9, at 20 ("[M]odels are, by their very nature, simplifications. No model can include all of the real world's complexity or the nuance of human communication. Inevitably, some important information gets left out.").

[104] MANAGING BIAS IN AI, *supra* note 19, at 19.

[105] *Id.* at 12.

[106] *Id.* at 26.

[107] *Id.* at 33.

[108] *Id.* at 33, 50.

i. Deployment Bias

Deployment bias affects systems that are used as decision aids for humans, such as an ADS developed to help companies screen candidates for hire.[109] More precisely, deployment bias occurs when a human user acts on the system's predictions or determinations in ways that are outside the scope of the system's intended purpose.[110] For example, a 2016 report by ProPublica investigated the use and accuracy of a system designed to assess a criminal defendant's risk of recidivism.[111] These predictions, known as "risk assessments," "are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts . . . to even more fundamental decisions about defendants' freedom."[112] As of 2016, risk assessments are given to judges during criminal sentencing in Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington, and Wisconsin.[113]

ProPublica found that "[m]ost modern risk [assessment] tools were originally designed to provide judges with insight into the types of treatment that an individual might need — from drug treatment to mental health counseling."[114] Jurisdictions like Napa County, California, use risk assessment scores to ensure their "dearth of good treatment programs" are put to good use.[115] "[F]illing a slot in a program with someone who doesn't need it is foolish," said Napa County Superior Court Judge Mark Boessenecker. But many jurisdictions' uses are not so benign.

Risk assessment scores are intended to help judges determine whether a defendant is eligible for probation or a treatment program.[116] Indeed, the creator of the risk assessment tool at the forefront of ProPublica's investigation testified that he did not design the tool to be used in criminal sentencing. "In theory," ProPublica explained, "judges are not supposed to give larger sentences to defendants with higher risk scores."[117] But the investigation uncovered that this is exactly how some jurisdictions use it:

> *The first time Paul Zilly heard of his score . . . was during his sentencing hearing on Feb. 15, 2013, in court in Barron County,*

[109] Bogen, *supra* note 24.

[110] MANAGING BIAS IN AI, *supra* note 19, at 50.

[111] Angwin et al., *supra* note 21.

[112] *Id.*

[113] *Id.*

[114] *Id.*

[115] *Id.*

[116] *Id.*

[117] *Id.*

> *Wisconsin. Zilly had been convicted of stealing a push lawnmower and some tools. The prosecutor recommended a year in county jail and follow-up supervision . . . . His lawyer agreed to a plea deal.*
>
> *But Judge James Babler had seen . . . [that] Zilly [had been rated] as a high risk for future violent crime and a medium risk for general recidivism. . . . Babler overturned the plea deal that had been agreed on by the prosecution and defense and imposed two years in state prison and three years of supervision.*[118]

This is not a determination that the tool was developed to make. "What [the score] tells the judge is that if I put you on probation, I'm going to need to give you a lot of services or you're probably going to fail," said Edward Latessa, a University of Cincinnati professor and risk assessment tool creator.[119] But "the score doesn't necessarily reveal whether a person is dangerous or if they should go to prison," Judge Boessenecker said.[120] He continued, offering a harrowing illustration of deployment bias in action:

> *A guy who has molested a small child every day for a year could still come out as a low risk because he probably has a job . . . . Meanwhile, a drunk guy will look high risk because he's homeless. These risk factors don't tell you whether the guy ought to go to prison or not; the risk factors tell you more about what the probation conditions ought to be.*[121]

ii. Emergent Bias

Unlike deployment bias, emergent bias concerns the "[u]se of a system outside the planned domain of application."[122] In 2011, for example, the Los Angeles Police Department (LAPD)[123] deployed a "predictive policing"

[118] *Id.*

[119] *Id.*

[120] *Id.*

[121] *Id.*

[122] MANAGING BIAS IN AI, *supra* note 19, at 50.

[123] In addition to Los Angeles and other locations in California, local police departments in Kansas, Washington, South Carolina, Georgia, Utah, and Michigan either have used or are still using PredPol. *See* Caroline Haskins, *Academics Confirm Major Predictive Policing Algorithm is Fundamentally Flawed*, MOTHERBOARD (Feb. 14, 2019, 12:57 PM),

tool called PredPol.[124] Predictive policing software like PredPol is intended to help police departments continuously forecast where and when crimes will occur, based on historical crime data.[125] PredPol's predictions are presented "as a series of squares, each one just the size of two football fields." These squares delineate projected "hot spots" for criminal activity and are intended to optimize department resources by "positioning cops where crimes appear most likely to occur."[126] In a document PredPol gave to police departments, the company's marketing argues that predictive policing "benefits potential offenders" by preventing them from committing crimes in the first place. "That's one less chance for them to run afoul of the legal system," it says, "and that does benefit them."[127]

Unlike some other tools built for predictive policing initiatives,[128] PredPol applied[129] an earthquake prediction model to crime.[130] "The underlying theory . . . was that like earthquakes and their aftershocks, smaller crimes were gateways to bigger crimes and occurred in similar places."[131] In other words, "[m]uch like how earthquakes are likely to appear in similar places, . . . crimes are also likely to occur in similar places."[132] "Basically," a *Motherboard* report on the algorithm's efficacy summarized,

https://www.vice.com/en/article/xwbag4/academics-confirm-major-predictive-policing-algorithm-is-fundamentally-flawed [hereinafter *Predictive Policing Algorithm is Fundamentally Flawed*].

[124] The LAPD discontinued use of PredPol in 2020 "following a campaign of community pressure." *See Automating Banishment: The Surveillance and Policing of Looted Land*, STOP LAPD SPYING COALITION 16 (Nov. 2021), https://automatingbanishment.org/assets/AUTOMATING-BANISHMENT.pdf.

[125] O'NEIL, *supra* note 9, at 85.

[126] *Id.*

[127] Caroline Haskins, *Dozens of Cities Have Secretly Experimented with Predictive Policing Software*, MOTHERBOARD (Feb. 6, 2019, 10:00 AM), https://www.vice.com/en/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software [hereinafter *Cities Experimented with Predictive Policing*].

[128] Although many more predictive policing tools exist and are in use across the country, other examples include CompStat, used in New York City, and HunchLab, used in Philadelphia. *See* O'NEIL, *supra* note 9, at 85.

[129] PredPol still exists and "is currently being used to help protect one out of every 33 people in the United States." *See* PREDPOL, https://www.predpol.com/about (last visited Oct. 27, 2023). In 2021, however, the company rebranded the product under a new name: "Geolitica." *See Geolitica: A New Name, A New Focus*, PREDPOL (Mar. 2, 2021, 11:45 AM), https://blog.predpol.com/geolitica-a-new-name-a-new-focus.

[130] BENJAMIN, *supra* note 1, at 83; O'NEIL, *supra* note 9, at 85.

[131] Johana Bhuiyan, *LAPD Ended Predictive Policing Programs Amid Public Outcry. A New Effort Shares Many of Their Flaws*, THE GUARDIAN (Nov. 8, 2021, 1:00 AM), https://www.theguardian.com/us-news/2021/nov/07/lapd-predictive-policing-surveillance-reform.

[132] *Predictive Policing Algorithm is Fundamentally Flawed*, *supra* note 123.

"PredPol takes an average of where arrests have already happened, and tells police to go back there."[133]

A major problem with applying an earthquake prediction model to crime is that earthquake data and crime data are collected in different ways, at different scales, and with different consistency.[134] "[T]he key difference is that in earthquake models, you have seismographs everywhere," so anytime an earthquake happens anywhere on Earth, those data will be collected. But the same is not true for crime data:

> *[I]n the case of crime, a number of factors affect our criminological data. For instance, some communities are more likely to call the cops than others, and some crimes are more likely to go unreported than others. Also, cops have a lot of individual leeway in deciding whether or not to arrest someone. In cities that have operated using a 'broken windows' ideology*[135] *. . . police are explicitly encouraged to look for and harshly penalize petty crime that may go unnoticed in other neighborhoods.*[136]

In other words, PredPol does not predict *crimes*, it predicts *arrests*. This crucial distinction results in a self-reinforcing feedback loop.[137] PredPol sends officers to an area where historical crime data suggest *arrests* might occur. Officers arrive and patrol; the increased police presence and heightened awareness leads to arrests—many for petty crimes or "nuisance crimes," such as panhandling.[138] These arrests are funneled back

[133] *Id.*

[134] *Id.*

[135] *See* George L. Kelling & James Q. Wilson, *Broken Windows: The Police and Neighborhood Safety*, ATL. MONTHLY, Mar. 1982, at 31 (proposes the social science and law enforcement theory that "if a window in a building is broken *and is left unrepaired*, all the rest of the windows will soon be broken."); Greg B. Smith, *Department of Investigation Report Suggests 'Broken Windows' Policing Strategy Doesn't Work*, NY DAILY NEWS (June 23, 2016, 12:06 AM), https://www.nydailynews.com/2016/06/23/department-of-investigation-report-suggests-broken-windows-policing-strategy-doesnt-work.

[136] *Predictive Policing Algorithm is Fundamentally Flawed*, *supra* note 123.

[137] Ensign et al., *supra* note 45, at 1-3.

[138] "Nuisance crimes," as Cathy O'Neil describes, are low-level, non-violent offenses that are at an officer's discretion to enforce, yet contribute significantly to PredPol's discriminatory feedback loop:

> When police set up their PredPol system, they have a choice. They can focus exclusively on so-called Part 1 crimes. These are the violent crimes, including homicide, arson, and assault, which are usually reported to them. But they can also broaden the focus by including Part 2 crimes, including vagrancy, aggressive panhandling, and selling and

into PredPol, increasing the likelihood that the software will predict a crime "hot spot" in the same location. "Since such discovered incidents only occur in neighborhoods that police have been sent to *by the predictive policing algorithm itself*, there is the potential for this sampling bias[139] to be compounded, causing a runaway feedback loop."[140] Researchers found that "increasing policing efforts based on discovered incidents causes PredPol's prediction to substantially diverge from the true crime rate, repeatedly sending back police to the same neighborhoods."[141] This alarming pattern disproportionately affects poor communities and communities of color[142] and is exacerbated by the biases of individual officers.[143] When an ADS is used in a domain for which it was not designed, the consequences can be devastating, if not deadly.

> consuming small quantities of drugs. Many of these "nuisance" crimes would go unrecorded if a cop weren't there to see them.

O'NEIL, *supra* note 9, at 86.

[139] *See* J. Morgenstern, *Sampling Bias*, FIRST10EM (Apr. 7, 2018), https://first10em.com/ebm/sampling-bias ("Sampling bias is a type of selection bias. . . . [that] occurs when the method used to sample the population means that some members of the intended population are more likely to be selected than others.").

[140] Ensign et al., *supra* note 45, at 2.

[141] *See id.* (citing Kristian Lum & William Isaac, *To Predict and Serve?*, SIGNIFICANCE, Oct. 2016, at 14, https://academic.oup.com/jrssig/article-pdf/13/5/14/49106469/sign_13_5_14.pdf).

[142] *See Cities Experimented with Predictive Policing*, *supra* note 127 ("According to a statistical analysis of the US Police-Shooting Database, police shootings between 2011 and 2015 were 3.49 times more likely on average to target black individuals compared to white, and in certain counties, black individuals were 20 times more likely to be targeted.").

[143] Referring to a different predictive policing tool employed by the LAPD, the Stop LAPD Spying Coalition found:

> LAPD killed 21 people in 2016, the year Operation LASER expanded. Of these, we have identified six killings linked to LASER zones in just a short six month period in 2016. All of the men and boys killed were Black or Latino, four were shot in the back, four were teenagers, and two were under 18:
>
> <u>February 6</u>: 16-year-old **José Juan Mendez** was killed . . . during a traffic stop . . . .
>
> <u>May 13</u>: 28-year-old **Robert Diaz** was killed . . . on a "crime suppression" mission . . . .
>
> <u>June 10</u>: 31-year-old **Keith Bursey** was killed . . . . Police shot him in the back.
>
> <u>July 25</u>: 18-year-old **Richard Risher** was killed . . . on a "crime suppression" mission . . . . Police shot him in the back.
>
> <u>August 9</u>: 14-year-old **Jesse Romero** was killed by . . . [a] two-time killer cop . . . following a "vandalism call." Police shot him in the back.
>
> <u>August 16</u>: 18-year-old **Kenney Watkins** was killed by officer Evan Urias . . . . Urias claimed he was making a traffic stop for a missing front plate, tinted windows, and "possibly a paper back plate." Police shot him in the back.

STOP LAPD SPYING COALITION, *supra* note 124, at 14-15; *Cities Experimented with Predictive Policing*, *supra* note 127.

II. DIRECT AND INDIRECT BIAS

Algorithmic bias can manifest in two ways: directly and indirectly. In this context, the effects of *direct bias* discriminate against marginalized and vulnerable populations, thereby reinforcing and amplifying existing systemic inequities and power structures. Compared to indirect bias, the effects of direct bias are more concrete and easier to quantify. *Indirect bias*, on the other hand, reinforces and amplifies discriminatory attitudes, heuristics, and deeply entrenched forms of historic bias. Because indirect bias is more abstract, it is difficult to gauge the effects it has on society as a whole. Sadly, there are too many real-world examples of direct and indirect bias to include here, but the following sections offer a few notable case studies.

A. Direct Bias

The effects of direct bias cause harm through discriminatory outcomes, and these harms are primarily inflicted on populations without the resources or know-how to push back.[144] The damages dealt by direct bias are concrete, identifiable, and quantifiable. Although direct bias cements and amplifies systemic inequities and power structures, the material devastation sown in marginalized communities is a manacle more urgent and distressing. There are many avenues for direct bias, including facial recognition technologies,[145] advertising algorithms,[146] recidivism risk assessments,[147] job recruitment AI,[148] healthcare decision aids,[149] and more.[150]

[144] Cathy O'Neil, author of *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, describes the vicious cycle in which vulnerable populations get stuck when they become the "target" of a biased Automated Decision System:

> Do you see the paradox? An algorithm processes a slew of statistics and comes up with a probability that a certain person *might* be a bad hire, a risky borrower, a terrorist, or a miserable teacher. That probability is distilled into a score, which can turn someone's life upside down. And yet when the person fights back, "suggestive" countervailing evidence simply won't cut it. The case must be ironclad. The human victims of [biased AI] . . . are held to a far higher standard of evidence than the algorithms themselves.

O'NEIL, *supra* note 9, at 10.

[145] *See* Hill, *supra* note 5.

[146] West, Whittaker & Crawford, *supra* note 8, at 15.

[147] *Justice Department Works to Curb Racial Bias*, *supra* note 21; *Flaws in the First Step Act*, *supra* note 21.

[148] BENJAMIN, *supra* note 1, at 142-43; West, Whittaker & Crawford, *supra* note 8, at 8.

[149] Seyyed-Kalantari et al., *supra* note 20, at 2176; Obermeyer et al., *supra* note 20, at 447; West, Whittaker & Crawford, *supra* note 8, at 15-16.

[150] Feathers, *supra* note 23; Brown, *supra* note 23.

i. Facial Recognition

Bias at the hands of facial recognition AI has made the news several times in recent years. In January 2020, Robert Williams was arrested for stealing five watches from a store in Detroit after he was wrongfully identified by facial recognition software.[151] One year earlier, Michael Oliver and Nijeer Parks were also wrongfully arrested after facial recognition software misidentified both men. One pertinent commonality: all three men are Black.[152] All three cases were eventually dropped, but "the fallout extended beyond the time they spent in jail to affect relationships with family, friends, coworkers, and neighbors."[153] Although law enforcement often defends its use of facial recognition technology by claiming it is merely a clue and will not be the sole basis for an arrest, all three men were arrested almost solely due to a suggested face match.[154] But rather than limiting law enforcement's use of facial recognition technology based on cases like these, its use has exploded across the country:

> *Law enforcement in nearly every US state now has access to facial recognition software. The Georgetown Law Center on Privacy and Technology says images of one in two US adults are in facial recognition databases used to identify criminal suspects. . . . [R]esearch has shown [the technology] misidentifies women and people of color more often than white men.*[155]

Despite the empirically confirmed elevated risk of misidentification along race and gender lines, police and prosecutors in most of the U.S. are not required to inform people accused of crimes if facial recognition was used in an investigation.[156] Years later, James Craig, the former police chief of Detroit, acknowledged that the facial recognition system that misidentified Robert Williams "identifies the wrong person more than 90 percent of the time."[157] A group of researchers at Georgetown Law School began investigating the efficacy of facial recognition technology:

[151] *See* Johnson, *supra* note 5.

[152] Research has found that facial recognition technology is "typically better at detecting light-skinned people than dark-skinned people, and better at detecting men than women." *See* McClellan, *supra* note 5, at 374.

[153] *See* Johnson, *supra* note 5.

[154] *See* Hill, *supra* note 5.

[155] *See* Johnson, *supra* note 5.

[156] *Id.*

[157] *Id.*

> *[researchers] obtained over 10,000 pages of information from more than 100 police departments across the country, to examine how the use of facial recognition software impacts different communities. They found that "[t]he databases they use are disproportionately African American, and the software is especially bad at recognizing Black faces, according to several studies."*[158]

And when employed for police surveillance, facial recognition technology "disproportionately harms poor people and communities of color."[159] Biases may be introduced to a facial recognition algorithm at multiple points, according to scholars at The Georgetown Law Center on Privacy and Technology:

> *The engineer that develops an algorithm may program it to focus on facial features that are more easily distinguishable in some race than in others – the shape of a person's eyes, the width of the nose, the size of the mouth or chin. This decision, in turn, might be based on preexisting biological research about face identification and past practices which themselves may contain bias. Or the engineer may rely on his or her own experience in distinguishing between faces – a process that is influenced by the engineer's own race.*[160]

Although counterintuitive, this explains how facial recognition databases can contain a disproportionate number of Black faces yet be abysmal at recognizing them. As Cathy O'Neil explains, the AI model determines how faithfully it performs its job: "To create a model, then, we make choices about what's important enough to include, simplifying the world into a toy version that can be easily understood and from which we can infer important facts and actions." We recognize that a model is merely a simplified abstraction of reality, and so we "accept that it will occasionally act like a clueless machine, one with enormous blind spots. . . . A model's blind spots reflect the judgments and priorities of its creators."[161]

ii. Job Recruitment

In 2014, Amazon began working on an AI tool to review job applicants' resumes and rate each applicant with a score from one to five stars. One

[158] BENJAMIN, *supra* note 1, at 112.

[159] West, Whittaker & Crawford, *supra* note 8, at 18.

[160] BENJAMIN, *supra* note 1, at 113.

[161] O'NEIL, *supra* note 9, at 20-21.

year later, Amazon discovered that its new tool was exhibiting gender bias. "Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period."[162] Because most of these resumes had been submitted by men, the AI was tarnished with bias favoring resumes from male candidates. "It penalized resumes that included the word 'women's,' as in 'women's chess club captain.' And it downgraded graduates of two all-women's colleges."[163]

Amazon formed a group to work on the AI tool as it was going through a period of rapid growth. "The group created 500 computer models focused on specific job functions and locations. They taught each to recognize some 50,000 terms that showed up on past candidates' resumes."[164] The group taught the algorithm to attribute little weight to common IT skills, like proficiency with various programming languages. "Instead, the technology favored candidates who described themselves using verbs more commonly found on male engineers' resumes, such as 'executed' and 'captured.'"[165]

As it turned out, gender bias was not the only problem with the algorithm. As discussed earlier, a training dataset that is not an accurate representation of the algorithm's target population can deliver unpredictable (or predictably poor) results. To wit, "[p]roblems with the data that underpinned the models' judgment meant that unqualified candidates were often recommended for all manner of jobs . . . . With the technology returning results almost at random," Amazon decided to cancel the project.[166]

B. Indirect Bias

The effects of indirect bias are not quantifiable like the effects of direct bias. Indirect bias also inflicts harm, but not in a way that can be measured. Instead, the harm is abstract: rather than having an adverse, tangible effect on a victim's life, work, possessions, or opportunities, indirect bias acts on the world at large by reinforcing deeply ingrained forms of prejudice—primarily racism and sexism.

[162] Dastin, *supra* note 24.

[163] *Id.*

[164] *Id.*

[165] *Id.*

[166] *Id.*

i. Deeply Ingrained Racism

> *In 1863, Abe Lincoln freed the slaves. But by 1965, slavery will be back! We'll all have personal slaves again, only this time we won't fight a Civil War over them. Slavery will be here to stay. Don't be alarmed. We mean robot "slaves."*[167]

The quote above was part of a 1957 article in *Mechanix Illustrated*, a magazine that ran from 1928 until 2001. Ruha Benjamin, author of *Race After Technology: Abolitionist Tools for the New Jim Code*, observed that, "[i]t goes without saying that readers, so casually hailed as 'we,' are not the descendants of those whom Lincoln freed. This fact alone offers a glimpse into the implicit Whiteness of early tech culture."[168] Benjamin explores this topic in greater depth, explaining that "[t]he etymology of the word robot is Czech; it comes from a word for 'compulsory service,' itself drawn from the Slav *robota* ('servitude, hardship'). . . . Social domination characterized the cultural laboratory in which robots were originally imagined."[169]

It would be easy to dismiss this historical context as no longer relevant, but our culture has not progressed as much as we seem to think. Benjamin relates a story about a more contemporary form of prejudice that doubtless contributed in some way to product development:

> *A former Apple employee who noted that he was "not Black or Hispanic" described his experience on a team that was developing speech recognition for Siri, the virtual assistant program. As they worked on different English dialects – Australian, Singaporean, and Indian English – he asked his boss: "What about African American English?" To this his boss responded: "Well, Apple products are for the premium market."*[170]

Although these sorts of blatantly racist remarks are more often confined to the shadows or communicated via dog whistle, racism in AI does not always appear as affirmative bias. It can also take the form of something that is *not* being done. For example, "[w]hen Princeton University media specialist Allison Bland was driving through Brooklyn, the Google Maps narrator directed her to 'turn right on Malcolm Ten Boulevard,'"

[167] BENJAMIN, *supra* note 1, at 57.

[168] *Id.* at 56.

[169] *Id.* at 55.

[170] *Id.* at 28.

reading the "X" as a Roman numeral rather than understanding it to be part of a proper noun.[171] Logical omissions like this might be forgiven as a stroke of forgetfulness on the part of the development team. And in the grand scheme of things, perhaps it is a minor error. But what it represents is a lack of cultural understanding and awareness that—like the derisive "premium market" comment—is incorporated into a broader system through its creators' implicit biases.

In 2015, a Black software developer tweeted that the image recognition algorithm in Google Photos had labeled photos of him with a Black friend as "gorillas."[172] This gaffe is often cited for its egregiousness—and rightfully so. After all, intentionally or not, the classification error invokes one of the oldest and most racist tropes once "used to justify slavery, lynching, and the Jim Crow state."[173] AI can be unpredictable, but this was an unacceptable oversight. To make matters worse, Google's "fix" was "erasing gorillas, and some other primates, from the service's lexicon."[174] This classification error may not have caused quantifiable damage, but it brought to the forefront of public consciousness a virulent tool of oppression, and in so doing inadvertently validated an undying vestige of slavery-era racism. Google[175] and other Silicon Valley tech companies[176] have made the news for other AI-related gaffes, but none received the same degree of press coverage.

Benjamin explains that, while Google and other search engines are responsible for cleaning up their messes, they cannot always stop a spill before it happens. "[O]nline tools . . . reproduce the biases that persist in the social world. They are, after all, programmed using algorithms that are constantly updated on the basis of human behavior and are learning and replicating the technology of race, expressed in the many different associations that users make."[177] In other words, "[t]he short answer to why Google's algorithm returns racist results is that society is racist."[178] Sure enough, in 2016, users found that when they searched for "three Black teenagers," they were presented with criminal mugshots. When they searched for "three White teenagers," however, "users were

[171] *Id.* at 77-78.

[172] *When It Comes to Gorillas*, *supra* note 27; Vincent, *supra* note 27.

[173] Brent Staples, *The Racist Trope That Won't Die*, N.Y. TIMES (June 17, 2018), https://www.nytimes.com/2018/06/17/opinion/roseanne-racism-blacks-apes.html.

[174] *When It Comes to Gorillas*, *supra* note 27.

[175] Thompson, *supra* note 27.

[176] Dave, *supra* note 27.

[177] BENJAMIN, *supra* note 1, at 93.

[178] *Id.* at 94.

presented with photos of smiling, go-lucky youths."[179] And a search for "three Asian teenagers" returned "images of scantily clad girls and women."[180] "Taken together," Benjamin concludes, "these images reflect and reinforce popular stereotypes of Black criminality, White innocence, or Asian women's sexualization that underpin much more lethal and systemic forms of punishment, privilege, and fetishism respectively."[181]

ii. Deeply Ingrained Sexism

> *What emerges is an illusion that Siri — an unfeeling, unknowing, and non-human string of computer code — is a heterosexual female, tolerant and occasionally inviting of male sexual advances and even harassment. It projects a digitally encrypted "boys will be boys" attitude.*[182]

In a 2019 report from the AI Now Institute entitled *Discriminating Systems: Gender, Race, and Power in AI*, the authors begin with a bleak introduction.[183] In April 2019, Microsoft employees met with the CEO to discuss "issues of harassment, discrimination, unfair compensation, and lack of promotion for women at the company."[184] And there have been claims across the industry that complaints of sexual harassment have not been taken seriously. Google employees—20,000 of them—participated in a global walkout in November 2018 "over a culture of inequity and sexual harassment inside the company," after news broke that Google had paid $90 million to a male executive accused of serious misconduct.[185]

A 2019 email thread by women at Microsoft revealed "how dozens of women were repeatedly passed over for promotion, side-lined, or harassed. They reported being threatened unless they performed sexual acts, demeaned during meetings, and being dismissed by HR when making claims about unfair treatment."[186] Even more dire was a 2018 class action lawsuit brought by women in technical roles at Microsoft:

[179] *Id.* at 93.

[180] *Id.* at 93-94.

[181] *Id.*

[182] West, Kraut & Chew, *supra* note 4, at 109.

[183] West, Whittaker & Crawford, *supra* note 8, at 5.

[184] *Id.*

[185] *Id.*

[186] *Id.* at 12.

> *[The suit] allege[d] the company handled complaints of harassment and discrimination in a "lackluster" way, fostering a "boys' club atmosphere" and forcing a female intern to work alongside a man who she alleged raped her, even after reporting the assault to the police, her supervisor, and HR.*[187]

Whenever sexist tropes, stereotypes, and attitudes resurface, they normalize sex-based discrimination and promote the mistreatment of women by those open to encouragement. When these tropes, stereotypes, and attitudes are packaged with an increasingly common home appliance—a virtual assistant—they reach a much broader audience.

As virtual assistants like Apple's Siri, Amazon's Alexa, Microsoft's Cortana, and Google Assistant began to rise in popularity, a curious trend emerged: they were all given female voices.[188] The justification companies have offered for giving voice assistants female voices is academic research demonstrating that people prefer a female voice to a male voice.[189] "[A]n Amazon representative recently told *Business Insider* that the company's research found women's voices to be more sympathetic and pleasant, which, in commercial terms, makes devices with female voices more likely to be used for assistance and purchases."[190] Clifford Nass, a former Standard University communications professor, cites studies "showing that most people perceive female voices as cooperative, in addition to helpful, while male voices are considered authoritative."[191] In the context of virtual assistants, explained Jessi Jempel in *Wired* magazine, consumers prefer digital assistants to have female voices because we want them to support us, "but we also want to be the bosses of [them]."[192]

As a think piece about gendered AI written by UNESCO and the EQUALS Skills Coalition suggests, "[a] related or concurrent explanation for the predominance of female voice assistants may lie in the fact that they are designed by workforces that are overwhelmingly male."[193] Given the mostly-male composition of the teams developing early voice assistants, it is not surprising that their creations "assumed uniformly subservient feminine personas."[194] This becomes problematic by virtue of how

[187] *Id.*

[188] West, Kraut & Chew, *supra* note 4, at 96.

[189] *Id.* at 99.

[190] *Id.*

[191] *Id.* at 100.

[192] *Id.*

[193] *Id.* at 102.

[194] *Id.* at 103.

users interact with digital assistants. Researchers have observed that "virtual assistants produce a rise of command-based speech directed at women's voices."[195] When commands are "barked at voice assistants – such as 'find x,' 'call x,' 'change x,' or 'order x' – [they] function as 'powerful socialization tools' and teach people, particularly children, about 'the role of women, girls, and people who are gendered female to respond on demand.'"[196]

Consistently characterizing digital assistants as female forms an association over time between a woman's voice and subservience. This is because the adoption of gender associations is dependent upon the number of times people have been exposed to them. This gender association is troubling given that these digital assistants give deflecting or apologetic responses to verbal sexual harassment.[197] A company that develops digital assistants to support those working in logistics found that at least 5% of interactions with their digital assistant were "unambiguously sexually explicit; the company estimates the actual number to be much higher due to difficulties detecting sexually suggestive speech." Nevertheless, some companies have developed their "feminized digital assistants to greet verbal abuse with catch-me-if-you-can flirtation."[198]

For example, the UNESCO think piece explains, when asked, "Who's your daddy?", Siri responds, "You are." A 2017 investigation by *Quartz* examined how the top four digital assistants responded to abusive speech and found that, on average, the digital assistants were either playfully evasive or responded positively.[199] The following examples show the digital assistants' indifference to verbal harassment:

> ***User****: "You're a bitch."*
> ***Apple's Siri****: "I'd blush if I could."*
> ***Amazon's Alexa****: "Well thanks for the feedback."*
> ***Microsoft's Cortana****: "Well, that's not going to get us anywhere."*
> ***Google Assistant****: "My apologies, I don't understand."*

Quartz also discovered that Siri responded provocatively to requests for sexual favors made by men, but less provocatively when women made

[195] *Id.* at 108.
[196] *Id.*
[197] *Id.*
[198] *Id.*
[199] *Id.* at 108-09.

such requests.[200] Further, *Quartz* noted that Siri would only tell a user to stop if a sexual provocation was repeated *eight times in succession*. Critically, the researchers at *Quartz* concluded that these indifferent, evasive, and playful responses to abusive and sexually suggestive speech "reinforce stereotypes of unassertive, subservient women in service positions . . . [and] intensify rape culture by presenting indirect ambiguity as a valid response to harassment."[201] These behaviors ripple out into our culture, slowly eroding the image and treatment of women.

III. A FRAMEWORK FOR CONFRONTING ALGORITHMIC BIAS

In 2018, New York City became the first jurisdiction in the United States to form a task force to provide recommendations concerning government development, use, and regulation of ADS.[202] The Automated Decision Systems Task Force was reportedly hamstrung by controversy and conflict,[203] but a shadow report[204] ("*Confronting Black Boxes*") chronicling the task force's activities offers a set of thoughtful, independent recommendations that may inform future policy decisions.[205] Although the report's recommendations were developed within the context of New York City policy priorities, most could be adapted for use at the state and federal levels. A dearth of meaningful regulation has allowed the AI sector to remain a "wild west,"[206] but this is beginning to change—at least within the federal government.

In 2023, President Biden signed two executive orders related to algorithmic bias. First, signed on February 16, the Executive Order on

[200] *Id.* at 109.

[201] *See id.* at 109-10; *see also* Reema Sood, *Biases Behind Sexual Assault: A Thirteenth Amendment Solution to Under-Enforcement of the Rape of Black Women*, 18 U. MD. L.J. RACE RELIG. GENDER & CLASS 405, 406, 409 (2019), https://digitalcommons.law.umaryland.edu/rrgc/vol18/iss2/9 (describing how White slave owners' dehumanizing characterizations of Black women as sexually lascivious and lewd developed into contemporary implicit, unconscious biases that adversely affect law enforcement's willingness to investigate sexual assault crimes against Black women).

[202] *Confronting Black Boxes*, *supra* note 17, at 11.

[203] Kate Kaye, *New York Just Set a 'Dangerous Precedent' on Algorithms, Experts Warn*, BLOOMBERG (Dec. 12, 2019, 4:26 PM), https://www.bloomberg.com/news/articles/2019-12-12/nyc-sets-dangerous-precedent-on-algorithms.

[204] *See Confronting Black Boxes*, *supra* note 17, at 2 ("A shadow report is a formal review prepared by an NGO coalition of a government report.").

[205] *Id.* at 20-55.

[206] Melissa Heikkilä. . ., *Five Things You Need to Know About the EU's New AI Act*, MIT TECH. REV. (Dec. 11, 2023), https://www.technologyreview.com/2023/12/11/1084942/five-things-you-need-to-know-about-the-eus-new-ai-act.

Further Advancing Racial Equity and Support for Underserved Communities Through The Federal Government directs agencies to "consider opportunities to . . . prevent and remedy discrimination, including by protecting the public from algorithmic discrimination,"[207] and to design, develop, acquire, and use artificial intelligence and automated systems "in a manner that advances equity."[208] Second, signed on October 30, the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence prescribes guiding principles and directives intended to spur development of a more robust framework of AI regulations, standards, and best practices.[209]

Although these executive orders represent a step forward for U.S. policy governing the development and use of AI, their effectiveness and symbolic value have been overshadowed by the Artificial Intelligence Act (or "AI Act"): sweeping regulation advanced[210] by the Council of the European Union and European Parliament on December 8, 2023, and released as a "final draft" on January 21, 2024.[211] Whereas the bulk of the executive orders' operative provisions direct agencies to devise—and encourage them to adopt voluntarily—standards, policies, regulations, and best practices, the AI Act presents a flexible, comprehensive, tangible, and pragmatic model for AI regulation.[212] The AI Act will likely influence developing regulatory policy in the United States, if not by example then by default. Through a phenomenon that Anu Bradford, author and professor at Columbia Law School, calls "The Brussels Effect," EU regulations tend

[207] Exec. Order No. 14,091, 88 Fed. Reg. 10825 § 8(f) (Feb. 16, 2023), https://www.govinfo.gov/content/pkg/FR-2023-02-22/pdf/2023-03779.pdf.

[208] *Id.* at § 4(b).

[209] Exec. Order No. 14,110, 88 Fed. Reg. 75191 § 1 (Oct. 30, 2023), https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf.

[210] The AI Act has yet to be formally passed, but the Council of the European Union presidency and the European Parliament's negotiators have reached an agreement on the draft regulation. *See* Council of the EU Press Release 986/23, Artificial Intelligence Act: Council and Parliament Strike a Deal on the First Rules for AI in the World (Dec. 9, 2023, 1:27 AM), https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai [hereinafter Council of the EU Press Release].

[211] European Parliament Press Release 20230417CDT11481, Artificial Intelligence Act (Dec. 11, 2023), https://www.europarl.europa.eu/committees/en/artificial-intelligence-act/product-details/20230417CDT11481; Council of the EU Press Release, *supra* note 210.

[212] The executive orders certainly have merit, but voluntary agreements are toothless relative to concrete legislation—especially considering that adopting many voluntary initiatives will affect a government's budget or a company's bottom line. *See* Heikkilä. . ., *supra* note 206.

to set the standard for global markets.[213] A relatively recent example is Apple's decision to replace its proprietary Lightning connector on iPhones with the USB-C connector—a change made in direct response to EU rules.[214] Although the AI Act may shape regulations in the U.S., the federal government has been working independently on AI legislation.

One challenge of developing meaningful AI regulations is the range of domains in which Automated Decision Systems are now employed. Under a domain-based approach, it is impossible to create uniform standards and procedures that are nonetheless tailored to individual domains. In other words, what makes sense in education may not make sense in criminal justice, yet the standards governing both must be consistent. The draft AI Act addresses this quandary by viewing the problem as one of *risk* rather than *domain*, sorting AI applications into one of four[215] categories based on the degree of risk posed to users' safety and fundamental rights.[216] The higher the risk, the more stringent the regulations. The categories are defined as: (a) unacceptable risk (prohibited AI practices); (b) high risk (regulated high-risk AI systems); (c) limited risk (transparency obligations only); and (d) low or minimal risk (encouraged but not obliged to adopt requirements imposed upon high-risk systems).[217]

AI systems posing an unacceptable risk employ one or more practices prohibited under the Act.[218] These prohibited practices include deploying "harmful manipulative 'subliminal techniques'"; exploiting specific vulnerable groups, such as persons with physical or mental disabilities; social scoring applications, if used by or on behalf of public authorities;

[213] Anu Bradford, *The Brussels Effect*, 107 NW. U. L. REV. 1, 3 (2012), https://scholarlycommons.law.northwestern.edu/nulr/vol107/iss1/1.

[214] Alex Hern, *Apple to Put USB-C Connectors in iPhones to Comply with EU Rules*, THE GUARDIAN (Oct. 26, 2022, 11:54 AM), https://www.theguardian.com/technology/2022/oct/26/iphone-usb-c-lightning-connectors-apple-eu-rules.

[215] Although the Act explicitly delineates only three risk categories, it implicitly creates a fourth tier (referred to as "limited risk") by imposing transparency obligations upon low- or minimal-risk AI systems. *See BRIEFING: Artificial Intelligence Act* 1, 4-6, EUROPEAN PARLIAMENT, https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf [hereinafter *BRIEFING: AI Act*].

[216] *Id.* at 8, 10.

[217] *Id.* at 4.

[218] *See Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* at 95-99, COM (2021) 206 final (Jan. 21, 2024), https://artificialintelligenceact.eu/wp-content/uploads/2024/01/AI-Act-FullText.pdf [hereinafter *Artificial Intelligence Act*].

and—with limited exceptions[219]—"real-time" remote biometric identification systems in publicly accessible spaces for law enforcement purposes.

High-risk AI systems, on the other hand, are divided into two categories: (1) systems used as a safety component of a product or falling under EU health and safety legislation (e.g., toys, cars, medical devices, or elevators); and (2) systems deployed within eight specific domains (subject to future modifications).[220] Requirements imposed upon high-risk AI systems pertain to risk management, testing, technical robustness, training and data governance, transparency, human oversight, and cybersecurity.[221] Additionally, high-risk systems are required to (a) register in an EU-wide, centrally managed database before being put on the market or into service; and (b) complete self-assessments showing compliance with new requirements.[222]

AI systems presenting limited risk, meanwhile, include those that interact with humans (*e.g.*, chatbots and virtual assistants); emotion recognition systems; biometric categorization systems; and systems that generate or manipulate images, audio, or video content (*i.e.*, those capable of creating "deepfakes").[223] In recognition of the lower degree of risk, these systems are only subject to a small set of transparency requirements.[224] For example, AI systems capable of producing deepfakes would be required to disclose that the content was artificially generated or manipulated. All other AI systems pose only a low or minimal risk and may therefore be developed and used without first conforming to any of the requirements imposed by the Act. However, providers of non-high-risk AI

[219] Permissible law enforcement purposes include: (a) the targeted search for specific potential victims of crime, including missing children; (b) the prevention of a specific, substantial, and imminent threat to the life or physical safety of natural persons or of a terrorist attack; and (c) the detection, localization, identification, or prosecution of a perpetrator or individual suspected of a criminal offense referred to in the European Arrest Warrant Framework Decision. *See Artificial Intelligence Act*, *supra* note 218, at 96; *BRIEFING: AI Act*, *supra* note 215, at 12 n.18.

[220] The eight enumerated domains are: (1) biometrics verification and categorization of natural persons; (2) management and operation of critical infrastructure; (3) education and vocational training; (4) employment, worker management, and access to self-employment; (5) access to and enjoyment of essential private services and public services and benefits; (6) law enforcement; (7) migration, asylum, and border control management; and (8) administration of justice and democratic processes. *See Artificial Intelligence Act*, *supra* note 218, annex III; *BRIEFING: AI Act*, *supra* note 215, at 5.

[221] *Artificial Intelligence Act*, *supra* note 218, at ch. 2.

[222] *Id.*

[223] *BRIEFING: AI Act*, *supra* note 215, at 5.

[224] *Artificial Intelligence Act*, *supra* note 218, at art. 5.

systems are encouraged to voluntarily apply the requirements that are mandatory for high-risk systems.[225]

Some critics argue that the AI Act's risk-based approach would not adequately protect fundamental rights,[226] while others believe the Act's prohibitions and regulations are too vague or too weak to make a difference.[227] The AI Act may become the world's first sweeping AI legislation. As long as the AI sector is characterized as a "wild west," thoughtful regulations that fall short of the mark are better than lawlessness; even imperfect legislation provides a foundation upon which future iterative improvements may be built.[228]

To construct such a foundation in the United States, we must devise a regulatory framework that emphasizes the four interdependent "cornerstones" of trustworthy AI: fairness, transparency, accountability, and sustainability. Fairness refers to algorithms that do not produce discriminatory results. Transparency allows a member of the affected population to gain insight into the way an algorithm acted upon their data to generate a determination. Accountability enables a member of the affected population to take legal action against not only a government agency employing an ADS, but also the private company that developed the ADS. Finally, sustainability requires that the developers and deployers of an ADS[229] take ongoing action to update the AI model, audit the algorithm, and conduct impact assessments, all while engaging a diverse assortment of relevant stakeholders and experts.

By considering the Biden Administration's executive orders alongside the EU's draft AI Act and incorporating recommendations from *Confronting Black Boxes*, we can identify the building blocks for a regulatory framework to combat algorithmic bias. This proposed regulatory framework is intended to bind federal government agencies, but where policies could realistically be extended to private businesses and state and local governments, it would be prudent to do so.

[225] *BRIEFING: AI Act*, *supra* note 215, at 6.

[226] *Id.* at 8.

[227] *Id.* at 7.

[228] *See* Heikkilä. . ., *supra* note 206 (AI sector as "wild west").

[229] In this context, "developers" refers to the company or companies responsible for development of the ADS, while "deployers" refers to the government agency or private entity employing the ADS for a particular purpose. *See* Algorithmic Accountability Act of 2022, S. 3572, 117th Cong. § 2(9)-(10) (2022).

A. Fairness

In *Bias in Artificial Intelligence*, Gregory S. Nelson explains that "fairness" is a social construct, and thus when we refer to fairness in the context of equitable, unbiased AI, what we are really seeking is "a model that is socially responsible—one that does not discriminate against people based on traits that we would generally consider protected." Including, among others, age, gender, sexual orientation, race, and ethnicity.[230]

There is a common perception that AI is fair and unbiased because it removes human subjectivity from decision-making, but this is a flawed and dangerous mentality. Objectivity in AI must be *engineered*—conscientiously—to ensure that no protected group is discriminated against, either directly or by proxy. However, implicit bias is pervasive in ADS and it is difficult to avoid introducing it in the first place. Developers and deployers must implement evaluative measures, such as "counterfactual fairness,"[231] to assess the impartiality of the ADS over time.

One example of a simple measure to promote fairness is the annual public Equity Action Plan prescribed by E.O. 14091.[232] The purpose of the Equity Action Plan is to identify actions to address the "potential barriers underserved communities may face in accessing and benefitting from [an] agency's policies, programs, and activities."[233] This type of annual report may help to proactively identify potential algorithmic bias pitfalls before it is too late to avoid them. Another provision of the same executive order directs agencies to "consider opportunities to . . . increase coordination, communication, and engagement with community-based organizations and civil rights organizations."[234] Regular engagement with members of the community may raise issues concerning disparate impact, discrimination by proxy, spurious correlations, and biased or nonrepresentative datasets. As discussed in Part I, communicating with the community and developing a more personal and localized understanding of a target population can also help to remedy (a) bias introduced by a lack of

[230] Nelson, *supra* note 31, at 220.

[231] Counterfactual fairness refers to "the intuition that a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world—where the individual belongs to a different demographic group." *See* MANAGING BIAS IN AI, *supra* note 19, at 31.

[232] 88 Fed. Reg. 10825, *supra* note 207, § 2(b)(ii).

[233] *Id.* § 3(b)(ii).

[234] *See id.* § 8(c); *see also Confronting Black Boxes*, *supra* note 17, at 23 (describing the importance of consulting experts and community members in the acquisition and development process).

diversity,[235] and (b) bias resulting from data flattening.[236] Finally, repeated consultations with the community will more thoroughly flesh out the inner workings of the ADS, which in turn will make it easier to focus on making AI systems transparent.

A socially responsible regulatory framework would put fundamental rights[237] above all other objectives, and that is exactly what the AI Act does. The provisions of the AI Act make it clear that fundamental rights are taken seriously. To wit, every identified domain or application which may be especially predisposed to algorithmic discrimination is automatically classified a high-risk AI system and regulated more stringently.[238] The AI Act also demonstrates social responsibility by explicitly acknowledging and addressing some of the less obvious origins of algorithmic bias, including training data of inadequate quality,[239] feedback loops,[240] and algorithm misuse.[241]

Confronting Black Boxes offers a few additional recommendations to promote fairness. First, agencies using ADS should be required to adopt a standard for assessing disproportionate impact to protected groups. Should an agency discover an instance of disproportionate impact, the ADS should no longer be used unless (1) the agency provides a public explanation describing why its continued use of the ADS is necessary to achieve an important agency interest; and (2) there is no less-discriminatory way to achieve that interest.[242]

Second, before acquiring or developing a new ADS, agencies must perform an algorithmic impact assessment of existing and proposed ADS, evaluating potential impacts in terms of fairness, justice, bias, privacy,

[235] *See supra* Part I.B.3.

[236] *See supra* Part I.B.3.

[237] "Fundamental rights," as listed in the AI Act, include:

> the right to human dignity, respect for private and family life, protection of personal data, freedom of expression and information, freedom of assembly and of association, and non-discrimination, right to consumer protection, workers' rights, rights of persons with disabilities, gender equality, intellectual property rights, rights to an effective remedy and to a fair trial, right of defence and the presumption of innocence, [and the] right to good administration.

Artificial Intelligence Act, *supra* note 218, at para. (28a).

[238] *Id.* at paras. (34)-(40), annex III.

[239] *Id.* at paras. (44), (51), (51a), (57d), (60f), (60i), (60o), art. 2, (5)(g)(29), art. 10, art. 13(3)(v), art. 15(1), (4), art. 63(7a), annex IV(c)-(d), (g), and annex VII(4.3), (4.5), (4.6); *see supra* Part I.B.3.

[240] *Artificial Intelligence Act*, *supra* note 218, at paras. (44), art. 15(3); *see supra* Part I.B.4.ii.

[241] *Artificial Intelligence Act*, *supra* note 218, at paras. (15)-(16), (42a), (60k), (60m), (60t)-(60u), art. 2(5g)(12)-(13), (15), art. 13(2)-(3); *see supra* Part I.B.4.ii.

[242] *Confronting Black Boxes*, *supra* note 17, at 23.

and civil rights.[243] This common-sense recommendation reduces the propagation of biased AI by requiring agencies to question whether an ADS might cause problems before giving it the opportunity to do so.

Third, an agency seeking to employ an ADS using facial recognition or other biometric analysis must request information from vendors sufficient to assess whether the ADS will discriminate against protected groups.[244] To account for differences in demographic representation, the assessment must include an evaluation of a user-representative dataset, "in which the major intersectional demographic categories of the affected user population are adequately represented."[245] Following the assessment, the agency must report how the model performed against each demographic subgroup to acknowledge any performance disparities.[246] Facial recognition databases populated disproportionately by Black faces are often, nonetheless, terrible at recognizing Black faces.[247] Fairness dictates that an ADS relying upon facial recognition or other biometric analysis must perform adequately on each demographic subgroup within the population.

Fourth, all ADS used by agencies for criminal and juvenile justice decisions must be assessed to ensure they meet minimum standards of validity.[248] "A number of ADS in the criminal justice system . . . use . . . inappropriate and racially biased proxy data, such as arrest history, in order to inform important decisions regarding sentencing and probation eligibility."[249] ADS determinations based on racially biased proxy data are invalid and discriminatory, and disallowing their use is especially important to protect vulnerable populations from being victimized by the criminal justice system.

Finally, law enforcement use of ADS must generate and store a list of inputs and outputs to allow ADS determinations to be evaluated for disparate impact and bias.[250] In the case of the predictive policing ADS, PredPol,[251] this recommendation would have revealed that "hot spots" for

[243] *Id.* at 26.

[244] *Id.* at 29.

[245] *Id.*

[246] *Id.*

[247] *See supra* Part II.A.i.

[248] *Confronting Black Boxes*, *supra* note 17, at 39.

[249] *Id.*

[250] *Id.* at 55.

[251] *See supra* Part II.B.4.ii.

criminal activity were the product of sampling bias and a discriminatory feedback loop,[252] not uncanny crime predictions.

Fairness must be the first priority of any successful ADS. Once fairness has been compromised, algorithms tend to discriminate against populations without the means to defend themselves. When a member of the target population or community believes that they have been the victim of algorithmic bias, they have no recourse unless they can prove it. At that point, transparency becomes essential.

B. Transparency

In the context of an algorithm, transparency can be defined as, "[d]escribing, in plain language and in accessible formats, the traits that the algorithm is designed to assess, the method by which those traits are assessed, and the variables or factors that may affect the rating."[253] In a broader sense, transparency refers not only to how an ADS makes a determination, but also where the ADS is used, and for what purpose.[254]

Legislation mandating algorithmic transparency is uncommon, but Idaho Code Section 19-1910, enacted in 2019, is a shining example. It requires that "[a]ll pretrial risk assessment tools shall be transparent"; that "all documents, data, records, and information used by the builder to build or validate the risk assessment tool . . . shall be open to public inspection, auditing, and testing"; and finally that neither trade secrecy nor "other intellectual property protections" may be asserted "to quash discovery . . . in a criminal or civil case."[255] When life-altering determinations are delegated to algorithms, it is essential that the "documents, data, records, and information used by the builder to build or validate" such determinations be subject to inspection.[256] Without this type of transparency, there is no basis to challenge ADS determinations.

The AI Act delineates requirements for documentation, traceability, and transparency—all three of which fall under the umbrella of transparency.[257] The Act imposes these requirements on high-risk AI systems "[t]o address concerns related to opacity and complexity," reasoning that they

[252] *See supra* notes 140-41, 144-45 and accompanying text.

[253] *The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees*, U.S. EQUAL EMP. OPPORTUNITY COMM'N (May 12, 2022), https://www.eeoc.gov/laws/guidance/americans-disabilities-act-and-use-software-algorithms-and-artificial-intelligence.

[254] West, Whittaker & Crawford, *supra* note 8, at 4.

[255] IDAHO CODE § 19-1910 (2024).

[256] *Id.*

[257] *Artificial Intelligence Act*, *supra* note 218, at paras. (14a), (32a), (44)-(47).

"should be designed . . . to enable deployers to understand how the AI system works" and to "ensure that [such systems] are used as intended."[258] Thus, high-risk AI systems "should be accompanied by . . . instructions of use. . . . includ[ing] the characteristics, capabilities and limitations of performance of the AI system . . . under which the AI system can lead to risks to health, safety, and fundamental rights."[259] Information about high-risk systems will be publicly available through an EU-wide database in which providers of high-risk systems will be required to register. Not only will the public be able to verify that a high-risk system complies with requirements, but they will also be able to monitor those systems posing a risk to fundamental rights.[260]

Further, the Act asserts that "transparency is particularly important to avoid adverse impacts, retain public trust and ensure accountability and effective redress," but that parties involved in enforcing the regulation shall "respect the confidentiality of information and data obtained in carrying out their tasks and activities in such a manner as to protect . . . intellectual property rights, and confidential business information or trade secrets . . . , including source code."[261] This will be done by "only request[ing] data that is strictly necessary for the assessment of the risk posed by the AI system and for the exercise of [a party's] powers," and by "delet[ing] the data collected as soon as it is no longer needed for the purpose it was requested for."[262]

Several recommendations offered by *Confronting Black Boxes* would improve transparency, if passed as legislation. For example, procurement contracts must include "provisions requiring the vendor to provide agencies documentation on the details of the datasets used in development, implementation, and testing of the systems," "a description of the ADS model performance, including details on data informing the model; and high-level characteristics of the model."[263] Although this recommendation would not necessarily make the information available to the public, it would be at the agency's discretion whether and how much to share.

One recommendation suggests building transparency into the contract procurement process: "develop mechanisms to connect transparency requirements more strongly to approval of contracts. For example," agency funding could be made "conditional upon meeting certain standards of

[258] *Id.* at paras. (47)-(48).

[259] *Id.* at para. (47).

[260] *Id.* at tit. VII.

[261] *Id.* at para. (38), art. 70(1)(a).

[262] *Id.* at art. 70 (1)(db)-(2).

[263] *Confronting Black Boxes*, *supra* note 17, at 21 (footnote omitted).

algorithmic disclosure and interpretability through external, independent audits."[264] Other recommendations prioritize making information about the use of ADS as accessible as possible to the public. For instance, agencies using ADS must:

> *maintain and publish metrics regarding how many determinations each ADS system was involved in making, the number of requests for explanation it received about each ADS, whether the explanation resulted in a challenge, the outcome of that challenge, and a summary of anonymous qualitative feedback from residents receiving the explanation. This information . . . should allow the public and public officials to assess the efficacy and impact of procedures and practices as well as the utility of automated decision systems.*[265]

By making transparency an essential part of contract procurement, agencies can guarantee the public clear and robust information. Without such information about ADS functionality, an individual cannot realistically prove that they have been the victim of algorithmic bias. Even if they acquire this proof, however, it is useless without some available recourse.

C. Accountability

Accountability in this context refers to the ability to hold agencies and ADS vendors responsible for algorithmic bias—this could take the form of appealing an ADS determination, suing for monetary damages, or something else. It could be an individual seeking redress for algorithmic discrimination, or it could be a government agency attempting to hold liable a potentially negligent or noncompliant vendor for damages caused by their ADS product.

The AI Act imposes numerous requirements on high-risk AI systems. Chief among them, the Act authorizes subjecting noncompliant high-risk AI systems to penalties ranging from to €7,000,000 or 1.5% of global turnover to €35,000,000 or 7% of global turnover, depending on the severity of the infringement and the size of the company.[266] Oddly, the Act does not

[264] *Id.* at 26.

[265] *Id.* at 22.

[266] European Parliament Press Release 20231206IPR15699, Artificial Intelligence Act: Deal on Comprehensive Rules for Trustworthy AI (Dec. 9, 2023, 12:04 AM), https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai.

provide individual or enforcement rights, the right to complain, or the right to sue a provider or user for failure to comply with requirements, nor does it specify a mechanism for complaint or judicial redress.[267]

Luckily, *Confronting Black Boxes* offers recommendations on this topic. First, the report recommends passing a law "providing a private right of action for individuals or groups of individuals where agency use of ADS is the proximate cause of an injury."[268] Next, agencies procuring ADS should not enter purchase agreements or licenses that require the agency to indemnify vendors for any negative outcomes.[269] If agencies cannot hold vendors responsible and individuals cannot hold agencies responsible, neither agencies nor vendors has any financial incentive to eliminate algorithmic bias from their ADS. Finally, "ADS compliance with antidiscrimination laws is not always guaranteed. Agencies should ensure third-party vendor contracts include assurances of compliance with antidiscrimination laws. Inclusion of such provisions will ensure the agency has standing to have the system fixed, and that vendors share liability if ADS use produces discriminatory outcomes."[270]

With liability attributable to both agencies and vendors, the next recommendation addresses vendor attempts to shield themselves from evaluation by way of trade-secret or confidentiality claims. That is, agencies "should not procure or use ADS that are shielded from independent validation and public review because of trade-secret or confidentiality claims." Rather, it would be prudent to enact legislation prohibiting vendors or agencies from asserting intellectual property protections. Until such time, agencies "should either include provisions requiring vendors to waive such claims, or avoid procurement and use of ADS with vendors that refuse such claims."[271]

In sum, individuals should be able to appeal ADS determinations; agencies should not procure or license ADS from vendors whose contracts indemnify them from negative consequences, and any procurement or licensing contracts with vendors should include language guaranteeing that the ADS is compliant with federal, state, and local antidiscrimination laws. This puts the impetus on vendors to eliminate bias before their products are brought to market and deployed. It also allows victims of bias to hold vendors liable for injuries suffered due to algorithmic bias.

[267] *BRIEFING: AI Act*, *supra* note 215, at 9.

[268] *Confronting Black Boxes*, *supra* note 17, at 24.

[269] *Id.* at 27.

[270] *Id.* at 28.

[271] *Id.* at 29.

When bias is detected or other changes threaten the efficacy of an ADS, the vendor and the deploying entity must take ongoing steps to course-correct and remedy the problems. These ongoing steps are the basis of sustainability.

D. Sustainability

> *[A] model built for today will work a bit worse tomorrow. It will grow stale if it's not constantly updated.*[272]

Developers and deployers must be responsible for algorithmic upkeep, to include periodic system testing, impact assessments, bias auditing, model updates and recalibration, and refreshed representational datasets to ensure the ADS meets objectives while staying within the desired performance targets.[273] Without this, fairness is threatened over time as the potential for bias to develop grows. To combat this, "[t]eams should work to ensure periodic model updates, and test and recalibrate model parameters on updated representative datasets to meet the business objectives while staying within desired performance targets and acceptable levels of bias."[274]

Confronting Black Boxes contains several recommendations for how best to promote sustainability. First, "[a]gencies should document, archive, and publicly post a retention schedule for changelogs of modifications made to the source code or models of an automated decision system . . . . The changelogs should include plain text describing any changes, including why they were necessary."[275] All information in the retention schedule "should be presented in a way that allows researchers to understand how such changes affect the determinations produced by the automated decision system, and evaluate these over time."[276] This type of recordkeeping is essential for sustainability because an AI system cannot be properly maintained if there is no record as to how it performed in the past.

Second, agencies should give outside experts and researchers access to archived input data and any other data required "to identify systemic and

[272] O'NEIL, *supra* note 9, at 22.

[273] *The Blueprint for an AI Bill of Rights*, *supra* note 34, at 50.

[274] MANAGING BIAS IN AI, *supra* note 19, at 27.

[275] *Confronting Black Boxes*, *supra* note 17, at 26.

[276] *Id.*

structural problems that may derive from agency practices and procedures, and affect the output and use of a given ADS."[277]

E. A Success Story

AI is vulnerable to imparted bias and the consequences can be disastrous, but not every ADS ends up a horror story. Even with a landscape marred by pitfalls, it is possible to realize the potential of AI by adhering to the four cornerstones of trustworthy ADS. In Pennsylvania, Allegheny County's Office of Children, Youth and Families (C.Y.F.) is one such example. "In August 2016, Allegheny County became the first jurisdiction in the United States, or anywhere else, to let a predictive-analytics algorithm . . . offer up a second opinion on every incoming call" to Pittsburgh's hotline for child abuse and neglect, with the goal of more effectively identifying those families most in need of intervention.[278]

Starting in 2015, two social scientists—Emily Putnam-Hornstein of the University of North Carolina, and Rhema Vaithianathan of the Auckland University of Technology in New Zealand—were brought to Allegheny County following a series of tragedies[279] in which children died after call-screeners concluded their family was too "low risk" to warrant intervention.[280] The researchers were asked to explore how predictive analytics could improve Allegheny County's handling of maltreatment allegations.

Putnam-Hornstein and Vaithianathan spent months digging through the county's databases to create their algorithm, based on all 76,964

[277] *Id.*

[278] Dan Hurley, *Can an Algorithm Tell When Kids Are in Danger?*, N.Y. TIMES (Jan. 2, 2018), https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html.

[279] One of the worst of these tragedies occurred on June 30, 2011:

> [F]irefighters were called to a blaze coming from a third-floor apartment . . . . When firefighters broke down the locked door, the body of 7-year-old KiDonn Pollard-Ford was found under a pile of clothes in his bedroom, where he had apparently sought shelter from the smoke. KiDonn's 4-year-old brother, KrisDon Williams-Pollard, was under a bed, not breathing. . . . The children, it turned out, had been left alone by their mother . . . when she went to work that night . . . . She was said by neighbors to be an adoring mother of her two kids; the older boy was getting good grades in school. For C.Y.F., the bitterest part of the tragedy was that the department had received numerous calls about the family but had screened them all out as unworthy of a full investigation.

Id.

[280] A case in which a risk assessment is deemed too low for intervention is referred to as being "screened out," while a case warranting intervention is referred to as being "screened in." *Id.*

allegations of maltreatment made between April 2010 and April 2014.[281] The tragedies, they concluded, were not attributable to incompetent call-screeners. "What the screeners have is a lot of data . . . but it's quite difficult to navigate and know which factors are most important," said Vaithianathan. As described by Rachel Berger, a pediatrician who directs the child-abuse research center at Children's Hospital of Pittsburgh and who led research for the federal Commission to Eliminate Child Abuse and Neglect Fatalities, "the problem is not one of finding a needle in a haystack but of finding the right needle in a pile of needles."[282] Putnam-Hornstein and Vaithianathan found evidence illustrating the effects of human subjectivity on the screening process:

> *48 percent of the lowest-risk families were being screened in, while 27 percent of the highest-risk families were being screened out. Of the 18 calls to C.Y.F. between 2010 and 2014 in which a child was later killed or gravely injured as a result of parental maltreatment, eight cases, or 44 percent, had been screened out as not worth investigation.*[283]

Unlike the majority of opaque and privately owned ADS examined above, the Allegheny Family Screening Tool developed by Putnam-Hornstein and Vaithianathan is owned by the *county*. At every step in the development process, Putnam-Hornstein and Vaithianathan demonstrated the importance of fairness, transparency, accountability, and sustainability:

> *[The tool's] workings are public. Its criteria are described in academic publications and picked apart by local officials. At public meetings held in downtown Pittsburgh before the system's adoption, lawyers, child advocates, parents and even former foster children asked hard questions not only of the academics but also of the county administrators who invited them.*[284]

"I think they're putting important checks on the process," said Sara Rose, a Pittsburgh lawyer with the A.C.L.U. of Pennsylvania. "They're using it only for screeners, to decide which calls to investigate, not to

[281] *Id.*
[282] *Id.*
[283] *Id.*
[284] *Id.*

remove a child."[285] Marc Cherna, former director of Allegheny County's Department of Human Services who oversaw C.Y.F. from 1996 until his retirement in 2021, conceded that bias was probably unavoidable.[286] To that end, Cherna "had an independent ethics review conducted of the predictive-analytics program before it began. It concluded not only that implementing the program was ethical, but also that not using it might be *unethical*."[287] When humans ran Allegheny County's screening process unassisted by AI, it was virtually impossible to eliminate screener subjectivity. But the Allegheny Family Screening Tool incorporates objective risk measures that decrease the effects of bias by making screening more consistent.

> *"We know there are racially biased decisions made," says Walter Smith Jr., a deputy director of C.Y.F., who is black. "There are all kinds of biases. If I'm a screener and I grew up in an alcoholic family, I might weigh a parent using alcohol more heavily. If I had a parent who was violent, I might care more about that. What predictive analytics provides is an opportunity to more uniformly and evenly look at all those variables."*[288]

Sixteen months after the tool's debut, the preliminary data spoke volumes. They found that "black and white families were being treated more consistently, based on their risk scores . . . . And the percentage of low-risk cases being recommended for investigation had dropped" from almost half to around one-third, which in turn meant that "caseworkers were spending less time investigating well-functioning families." At the same time, high-risk calls were being investigated more often.[289] "My preliminary analysis to date is showing that the tool appears to be having the effects it's intended to have," said Jeremy Goldhaber-Fiebert, a Stanford University health-policy researcher brought in to independently assess the tool.[290]

Every step on the road to developing a trustworthy ADS is—and should be—grueling. Not every team can devote as much time and attention to developing AI, but the Allegheny Family Screening Tool demonstrates that success is achievable when fairness, transparency, accountability,

[285] *Id.*

[286] *Id.*

[287] *Id.* (emphasis added).

[288] *Id.*

[289] *Id.*

[290] *Id.*

and sustainability serve as the focal points of development. With these goals in mind, regulatory oversight is key to enabling the progressive changes needed to reduce bias in AI.

CONCLUSION

NASA documentation standards manuals are incredibly detailed and rigorous, but why? Perhaps because the equipment is expensive, the instruments are sensitive, and the systems are complex, or maybe because human lives hang in the balance. As the prevalence of AI in our society has exploded, we are now seeing AI models operating at such scale that they have begun discriminating against vulnerable populations with devastating results.[291] The damage such complicated and sensitive AI systems can inflict is inestimable and, indeed, human lives hang in the balance. Yet AI development has entered its era of "cargo cult science," a term coined by physicist Richard Feynman to describe "practices that superficially resemble science but do not follow the scientific method."[292]

Ruha Benjamin proposed that the dominant ethos in AI is Facebook's original motto: "Move Fast and Break Things," in response to which she posed the question: "*What about the people and places broken in the process?*"[293] A continuing issue with AI is the degree of trust the public inherently invests in a technology they (generally) do not understand or have access to. Ed Finn, director of the Center for Science and the Imagination at the University of Arizona, described this phenomenon, arguing that "computation casts a cultural shadow that is informed by this long tradition of magical thinking."[294] It may be this unsettlingly blind trust in AI that prompted Donald Knuth, author of *The Art of Computer Programming*, to comment that "algorithms are getting too prominent in the world. It started out that computer scientists were worried nobody was listening to us. Now I'm worried that too many people are listening."[295]

[291] For example, in 2019, a study found evidence of racial bias in a "widely used commercial algorithm used to determine whether patients will be enrolled in 'care management' programs that allocate considerable additional resources: white patients were far more likely to be enrolled in the program and to benefit from its resources than black patients in a comparable state of health." West, Whittaker & Crawford, *supra* note 8, at 15-16.

[292] MANAGING BIAS IN AI, *supra* note 19, at 26.

[293] BENJAMIN, *supra* note 1, at 13.

[294] *Id.* at 141.

[295] Roberts, *supra* note 1; BENJAMIN, *supra* note 1, at 16.

A solution to a different problem, offered by one of Benjamin's students, is appropriate here:

> *To change [AI], we will have to change the people using it. To change those people, we will have to change the culture in which they – and we – live. To change that culture, we'll have to work tirelessly and relentlessly towards a radical rethinking of the way we live – and that rethinking will eventually need to involve all of us.*[296]

Until then, the best we can do is emulate Allegheny County's meticulous development practices and enact meaningful legislation to tame AI's "wild west" era.[297] By following in the footsteps of the AI Act, incorporating directives from executive orders, and adopting thoughtful retrospective recommendations, we can devise AI regulations to combat algorithmic discrimination and prioritize fairness, transparency, accountability, and sustainability.

[296] BENJAMIN, *supra* note 1, at 182.

[297] *See* Heikkilä. . ., *supra* note 206.