United States of America
FEDERAL TRADE COMMISSION
WASHINGTON, DC 20580

Remarks of Commissioner Rebecca Kelly Slaughter

Algorithms and Economic Justice

UCLA School of Law
January 24, 2020

I. Introduction

Good afternoon, my name is Rebecca Kelly Slaughter,[1] and it is my pleasure to be here with you this afternoon to discuss the potentially transformative power of AI-driven algorithms—and how best to use this power to promote justice and expand opportunity. Research, recent examples, and today's rich discussions highlight the perils that flawed algorithmic decision-making can have in the area of criminal justice: over-surveillance, over-policing, wrongful detainment and arrest, and biased risk assessments used to determine pre-trial status and even sentencing.[2] Our criminal justice system has long struggled to deliver equitable justice, and we must act to prevent the entrenchment of algorithmic decision-making tools that produce the same biased outcomes—or worse—that we are striving to reduce.

The dangers of flawed algorithms are not limited, however, to criminal justice. I have the honor of serving as a Commissioner at the Federal Trade Commission, which is a civil law enforcement agency. The FTC's mission is to protect consumers from unfair and deceptive practices and to promote competition in the marketplace. In other words, our primary focus is on economic and civil justice rather than criminal justice.

The cornerstone of economic justice is the protection of equal opportunities. Algorithmic decision-making has the potential to further economic justice by distributing opportunity more broadly, resources more efficiently, and benefits more effectively. Pairing dramatically deeper pools of data with rapidly advancing machine-learning technology offers a chance at substantial benefits for consumers. When used successfully, AI has, for example, transformed access to educational opportunities[3] and improved health outcomes through improved diagnostic rates and

[1] The views expressed in these remarks are my own and do not necessarily reflect the views of the Federal Trade Commission or any other commissioner.

[2] Elec. Privacy Info. Ctr., "Algorithms in the Criminal Justice System: Pre-Trial Risk Assessment Tools," *EPIC.org*, https://epic.org/algorithmic-transparency/crim-justice (last visited Jan. 17, 2020).

[3] Matt Kasman & Jon Valant, "The Opportunities and Risks of K-12 Student Placement Algorithms," *Brookings* (Feb. 28, 2019), https://www.brookings.edu/research/the-opportunities-and-risks-of-k-12-student-placement-algorithms/.

care adjustments.[4]

One of the most compelling uses for AI-powered algorithms is to eliminate the biases that infect human decision-making. Terms such as machine-learning, math, code, and data hold out the tantalizing prospect of objective, unbiased, and superior decision-making. Some of this promise bears out. But we also know that algorithmic decisions still produce biased and discriminatory outcomes. We have seen mounting evidence of AI-generated economic harms in employment, credit, healthcare, and housing.

These harms tend to fall into three broad categories: denial of a benefit**,** meaning you don't get the job, the loan, the house, or the healthcare; exclusion from opportunity, meaning you don't even *see* the job posting, the refinance offer, or the home for sale; and negative or predatory targeting, meaning you're hit with higher prices or worse terms. Companies may not set out intending to discriminate based on a consumer's race or gender, yet these discriminatory outcomes continue.

In my role as an FTC Commissioner, I have the opportunity and the obligation to consider many different arguments and justifications for why particular practices are lawful and promising, or, alternatively, why they are illegal and dangerous. I endeavor to listen carefully with a thoughtful and critical ear to all of these arguments and often find many of them persuasive. But I have observed that, within the realms of the technological fields in which we conduct enforcement, there is a tendency to name a particular technology as though the technology itself is the entirety of the explanation. This is what I think of as the "Because AI" phenomenon: the argument that a particular practice is either beyond scrutiny or beyond redemption "because it is AI." This argument is pretty much never persuasive to me.

(For what it's worth, this is not a problem limited to AI. I also object to "because blockchain," "because 5G," and "because China." But it is egregiously common in the AI field.)

The "because AI" phenomenon is a manifestation of the temptation to see this new technology as either a panacea for the world's ills or a pandemic that can only exacerbate them. It is neither. R. David Edelman, who is currently at MIT, has an expression: "AI is not magic; it is math and code." I think that is exactly the right adage to keep in mind. As we consider the threats that algorithms pose to justice—whether criminal or economic—we must remember that, just as the technology is not magic, neither are the cures to its shortcomings.

The problems posed by AI are both nuanced and context-specific. And because many of the flaws of algorithmic decision-making have long-standing human decision-making analogs, we have a body of enforcement experience from which we can and should draw. These lessons from civil enforcement may resonate with those of you who are more focused on the criminal-justice side, because many of the factors that contribute to flawed algorithmic decision-making in the marketplace are similar to those that plague algorithms in the criminal-justice system.

[4] Irene Dankwa-Mullan, et al., "Transforming Diabetes Care Through Artificial Intelligence: The Future is Here," 22 *Popular Health Mgmt.* 229 (2019), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6555175.

I want to use my time to share my civil-enforcement perspective on several issues that have been woven through today's discussions: First, what causes algorithms to produce discriminatory outcomes or other harms to consumers? Second, what steps are most effective at mitigating these harms and improving algorithmic outcomes? Third, what can we do under our current law to advance these improvements? Finally, what could new rules or regulations offer in terms of increased protection from algorithmic harms?

II. Causes of Problematic Algorithmic Outcomes

When we focus on the most troubling examples of flawed algorithms in the marketplace in recent years, there emerges a clear list of factors that contribute to discriminatory or unsavory outcomes: faulty inputs, faulty conclusions, a failure to adequately test, and proxy discrimination. In many cases, these four factors work in concert, but I'd like to begin by spending a few minutes on each of them individually.

Faulty Inputs

A machine-learning algorithm can only be as useful as the data used to develop it, and faulty inputs can produce thoroughly problematic outcomes. This broad concept is captured in the familiar phrase "garbage in, garbage out."

The data used to develop a machine-learning algorithm might be skewed—either because individual data points reflect problematic human biases, or because the overall dataset is not adequately representative, or a combination of these factors. Often this skewed training data reflects historical and enduring patterns of prejudice or inequality; these faulty inputs can create biased algorithms that exacerbate these societal injustices.

One example that comes to mind is Amazon's failed attempt to develop a hiring algorithm driven by machine learning. As Reuters reported in late 2018, that effort was ultimately abandoned prior to deployment because the algorithm would have systematically discriminated against women. This issue stemmed from the fact that the resumes used to train Amazon's algorithm reflected the male-heavy skew in the company's applicant pool, and, despite their engineers' best efforts, the algorithm kept identifying this pattern and attempting to reproduce it.[5]

Faulty Conclusions

A different type of problem involves the use of data to generate conclusions that are inaccurate or misleading—perhaps better phrased as "data in, garbage out." This type of algorithmic flaw, faulty conclusions, forms the basis for much of the rapidly proliferating field of AI-driven "affect recognition" technology. Many companies claim that their affect recognition products can accurately detect an individual's emotional state by analyzing her facial

[5] Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," *Reuters* (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

expressions, eye movements, tone of voice, or even her gait.[6]

The underlying algorithms are designed to find patterns in, and reach conclusions based upon, certain types of physical presentations and mannerisms. But, as one might expect, human character cannot be reduced to one or two observable factors. For example, consider the algorithmic analysis of facial expressions—one popular flavor of affect-recognition technology. A major psychological review published last summer analyzed over a thousand studies on emotional expression and concluded that "[e]fforts to simply 'read out' people's internal states from an analysis of their facial movements alone, without considering various aspects of context, are at best incomplete and at worst entirely lack validity, no matter how sophisticated the computational algorithms."[7] Nevertheless, companies such as Microsoft,[8] IBM,[9] and Amazon[10]—as well as a host of well-funded start-ups—continue to sell this questionable technology, and it is sometimes deployed to afford or deny people formative life opportunities.[11]

A striking example of the use of affect-recognition technology is in hiring. A number of companies claim their products are capable of reliably extrapolating personality traits and predicting social outcomes such as job performance.[12] Their methods of "analysis" often involve

[6] *See* Kate Crawford et al., *AI Now 2019 Report* 50-52 (2019), https://ainowinstitute.org/AI_Now_2019_Report.pdf; Manish Raghavan et al., "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices," 12 (arXiv: 1906.09208, 2019), https://arxiv.org/abs/1906.09208.

[7] Lisa Feldman Barrett et al., "Emotional Expressions Reconsidered Challenges to Inferring Emotion from Human Facial Movements," 20 *Psychol. Sci. Pub. Interest* 48 (2019); *id.* at 1, 46-51 (explaining that "how people communicate anger, disgust, fear, happiness, sadness, and surprise varies substantially across cultures, situations, and even people within a single situation. Furthermore . . . a given configuration of facial movements, such as a scowl, often communicates something other than an emotional state"); *see also* Zhimin Chen & David Whitney, "Tracking the Affective State of Unseen Persons," *Psychol. Cognitive Sci.* (Feb. 5, 2019), https://www.pnas.org/content/pnas/early/2019/02/26/1812250116.full.pdf (finding that detecting emotions with accuracy requires more information than is available just on the face and body).

[8] In late 2018, one researcher ran Microsoft's Face API on a public dataset of NBA player pictures and found that it interpreted Black players as having more negative emotions than White players. *See* Laruen Rhue, "Racial Influence on Automated Perceptions of Emotions," (last revised Dec. 17, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281765.

[9] Taylor Telford, "'Emotion Detection' AI is a $20 Billion Industry. New Research Says It Can't Do What It Claims," *Washington Post* (July 31, 2019, 1:27 PM), https://www.washingtonpost.com/business/2019/07/31/emotion-detection-ai-is-billion-industry-new-research-says-it-cant-do-what-it-claims.

[10] Amazon recently claimed that its Rekognition service can now identify fear. Saheli Roy Choudhury, *Amazon Says Its Facial Recognition Can Now Identify Fear*, CNBC (Aug. 14, 2019, 11:58 AM), https://www.cnbc.com/2019/08/14/amazon-says-its-facial-recognition-can-now-identify-fear.html.

[11] Today, the affect recognition industry is worth $20 billion, and some analysis see it exploding to $90 billion by 2024. *See* Telford, *supra* note 9; Paul Sawers, "Realeyes Raises $12.4 Million to Help Brands Detect Emotion Using AI on Facial Expressions," *Venture Beat* (June 6. 2019, 12:30 AM), https://venturebeat.com/2019/06/06/realeyes-raises-12-4-million-to-help-brands-detect-emotion-using-ai-on-facial-expressions/; "Emotion Detection and Recognition (EDR) Market – Growth, Trends, and Forecast (2020-2025)," https://www.mordorintelligence.com/industry-reports/emotion-detection-and-recognition-edr-market (last visited, Jan. 22, 2020).

[12] Rebecca Heilweil, "Artificial Intelligence Will Help Determine If You Get Your Next Job," *Recode* (Dec. 12, 2019), https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen (providing examples

questionable assessments of observable physical factors.[13] Such algorithmic hiring products merit skepticism in any application, and recent studies suggest they might systematically disadvantage applicants with disabilities because they present differently than the majority of a company's applicants or employees.[14] These reports should trouble any employer using an AI hiring product to screen applicants.

Closer to the "digital snake oil" side of the spectrum, a recent study of AI-driven employment screening products highlights one company that purports to profile over sixty personality traits relevant to job performance—from "resourceful" to "adventurous" to "cultured"—all based on an algorithm's analysis of an applicant's 30-second recorded video cover letter.[15]

Pseudo-science claims of power to make objective assessments of human character are not new; consider handwriting analysis that purports to reveal one's personality or even the lie-detector polygraph testing that has long been inadmissible in court. Despite the veneer of objectivity that comes from throwing around terms such as "AI" and "machine learning," the technology is still deeply imperfect.[16] And in this way, "AI-powered" claims can be more pernicious than their analog counterparts because they can engender less skepticism. [17]

of companies that use AI in recruiting (Arya and Leoforce), initial contact with a potential recruit or reconnecting a prior candidate (Mya), personality assessments (Pymetrics), and video interviews (HireVue)).

[13] Arvind Narayanan, an Associate Professor of Computer Science at Princeton University and leader of the Princeton Web Transparency and Accountability Project, has criticized claims that AI job assessments can evaluate speech patterns and body language, describing these products as "fundamentally dubious." *See* Arvind Narayanan, Presentation: How to Recognize AI Snake Oil, https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.

[14] *See* Anhong Guo et al., "Toward Fairness in AI For People With Disabilities: A Research Roadmap," 4 (arXiv: 1907.02227, 2019), https://arxiv.org/abs/1907.02227; Jim Fruchterman & Joan Melllea, *Expanding Employment Success for People with Disabilities* 3 (2018), https://benetech.org/wp-content/uploads/2018/11/Tech-and-Disability-Employment-Report-November-2018.pdf.

[15] *See* Manish Raghavan et al., *surpa* note 6, at 11. Of course, not all AI hiring algorithms are this potentially problematic, but evidence suggests that other products can suffer from similar structural shortcomings. *Id.* at 513.

[16] Algorithmic hiring is problematic for a number of other reasons. For example, we're already seeing the development of a market for strategies and products that are designed to 'beat' different kinds of hiring algorithms. Some people will be unable to afford these services, and they will be judged against those who can, creating another barrier to employment that perpetuates historical wealth inequality and hinders social mobility. *See* Sangmi Cha, "'Smile with Your Eyes': How to Beat South Korea's AI Hiring Bots and Land a Job, *Reuters* (Jan. 12, 2020, 8:15 PM), https://www.reuters.com/article/us-southkorea-artificial-intelligence-jo/smile-with-your-eyes-how-to-beat-south-koreas-ai-hiring-bots-and-land-a-job-idUSKBN1ZC022; Hilke Schellmann, "How Job Interviews Will Transform in the Next Decade," *Wall St. J.* (Jan. 7, 2020, 9:48 AM), https://www.wsj.com/articles/how-job-interviews-will-transform-in-the-next-decade-11578409136.

[17] Additionally, many companies capitalize on the positive associations with AI, despite the fact that they do not even *use* AI in any material way for their business. One recent report found that a full 40% of European startups that were classified as AI companies do not accurately fit that description, and startups with the AI label attract 15% to 50% more in their funding rounds than other technology startups. *See* Parmy Olson, "Nearly Half of All 'AI Startups' are Cashing In on Hype," *Forbes* (Mar. 4, 2019), https://www.forbes.com/sites/parmyolson/2019/03/04/nearly-half-of-all-ai-startups-are-cashing-in-on-hype/#151cd215d022.

At the end of the day, an employment-screening algorithm's assessment of a candidate can actually be *less* accurate than the subjective impression I get when I conduct an interview. These risks can be compounded when certain products are emphatically marketed as producing reliable predictions about potential hires when these conclusions are in fact flawed and misleading. This is a phenomenon with which we are quite familiar at the FTC: new technology, same old lack of substantiation.

Failure to Test

Even when an algorithm is designed with care and good intentions, it can still produce biased or harmful outcomes that are not anticipated. Too often, algorithms are deployed without adequate testing that could uncover these unanticipated outcomes before they harm people in the real world. And, as we frequently caution in the area of data security, while pre-deployment testing is an important step, it is not sufficient to prevent problems.[18] Constant monitoring, evaluating, and retraining are critical to identify and correct unintentional bias and disparate outcomes.

The healthcare field provides good examples of bias that can result from a failure to adequately assess the variables used in an algorithm pre-deployment *and* a failure to monitor outcomes and test for bias post-deployment. A recent study published in *Science* found racial bias in a widely used machine-learning algorithm intended to improve access to care for high-risk patients with chronic health problems.[19] The algorithm used health care costs as a proxy for health needs, but, for a variety of reasons unrelated to health needs, White patients spend more on health care than their equally sick Black counterparts. Considering health care costs as the indicator of need therefore caused the algorithm to disproportionately flag White patients for additional care.[20] As a result of this imbedded bias, researchers estimated that the number of Black patients identified for extra care was reduced by more than half. The potential scale of this harm is staggering: the researchers called this particular healthcare algorithm "one of the largest and most typical examples of a class of commercial risk-prediction tools that . . . are applied to roughly 200 million people in the United States each year."[21]

[18] An additional caution on the subject of data security: the consumer protection challenges posed by algorithmic decision-making are not limited to those discussed here; the vast quantities of information involved in these algorithms can lead to serious concerns about storage, security, and proper disposal of data. Companies that get the security side of the equation wrong may be violating data security rules as well as causing the more AI-specific harms discussed in this speech.

[19] Ziad Obermeyer et al., "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," 366 *Science* 447 (2019), https://science.sciencemag.org/content/366/6464/447.

[20] Sujata Gupta, "Bias in a Common Health Care Algorithm Disproportionately Hurts Black Patients," *ScienceNews.org* (Oct. 24, 2019, 2:00 PM), https://www.sciencenews.org/article/bias-common-health-care-algorithm-hurts-black-patients ("because of the bias. . . healthier white patients get to cut in line ahead of black patients, even though those black patients go on to be sicker.").

[21] The researchers continue: "It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health care sector . . . [an industry] in which algorithms are already used at scale today, unbeknownst to many." Ziad Obermeyer et al., *supra* note 19, at 447.

The researchers who uncovered the flaw in the algorithm were able to do so because they looked beyond the algorithm itself to the outcomes it produced and because they had access to enough data to conduct a meaningful inquiry.[22] It is also worth noting that when the researchers identified the flaw, the algorithm's manufacturer worked with them to mitigate its impact, ultimately reducing bias by 84%.[23] But a comprehensive inquiry into the potential limitations of using costs as a proxy for sickness, including relevant social context, should have raised concerns pre-deployment.[24] And while there is no simple test to reliably detect and prevent bias, early and ongoing testing of the outcomes in this instance may have caught this flaw years earlier.

Another recent example spotlighting potential algorithmic bias, anecdotal this time, involved the Apple Card. This fall, a noted software programmer, David Heinemeier Hansson, took to Twitter to decry the fact that he had been given an Apple credit line 20 times higher than his wife's, despite the fact that they had shared finances and her credit score was higher.[25] Many others joined in the discussion to report a similar experience in their relative marital Apple Card credit limits, including Steve Wozniak. The best explanation Hansson reported being able to get was that this was not discrimination—it was "just the algorithm."[26] We don't know—yet—what caused these disparities. But the stories suggest that, at a minimum, further testing and better understanding of the algorithm driving these credit outcomes would have been beneficial.[27]

One more example on this front: A few years ago a reporter found that when you typed in a number of common female names on LinkedIn, you would be prompted with a similarly spelled man's name instead—Stephan Williams when you searched for Stephanie Williams.[28] But, according to the reporter, when you typed in any of the 100 most common male names—LinkedIn *never* prompted you with a female alternative; still the company denied there was any

[22] The researchers analyzed data on patients at one hospital that used the high-risk care algorithm and focused on 40,000 patients who self-identified as White and 6,000 who identified as Black during a two-year period. The algorithm had given all patients a risk score based on past health care costs. In theory, patients with the same risk scores should be similarly sick. Instead, on average Black patients with the same risk scores as White patients had more chronic diseases. Gupta, *supra* note 20.

[23] Ziad Obermeyer et al., *supra* note 19.

[24] *See generally*, Nicole Wetsman, "There's No Quick Fix to Find Racial Bias in Health Care Algorithms," *The Verge* (Dec. 4, 2019, 10:36 AM) https://www.theverge.com/2019/12/4/20995178/racial-bias-health-care-algorithms-cory-booker-senator-wyden ("Algorithms that use proxy measures, for example—like health costs as a measure of sickness—need to be examined more carefully, he says, and any bias in the proxy would have to be evaluated separately.").

[25] Alisha Haridasani Gupta, "Are Algorithms Sexist?" *N.Y. Times* (Nov. 15, 2019), https://www.nytimes.com/2019/11/15/us/apple-card-goldman-sachs.html.

[26] Hansson recounted escalating levels of Apple customer service representatives being unable to explain the discrepancy, one even stating that Apple was not discriminating: "IT'S JUST THE ALGORITHM." @dhh, *Twitter*, (Nov. 7, 2019, 12:34 PM), https://twitter.com/dhh/status/1192540900393705474?s=20.

[27] Gupta, *supra* note 25.

[28] Matt Day, "How LinkedIn's Search Engine May Reflect a Gender Bias," *Seattle Times*, Sept. 8, 2016, https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias.

algorithmic bias.[29] Here again we have an example of a potentially biased outcome, uncovered through user testing, that might have been prevented altogether if the platform engaged in regular outcome testing.

It is entirely possible that these examples were a product of both faulty inputs *and* a failure to test—algorithmic bias will often be the product of multiple flaws. But what stands out to me about each of these examples is that additional testing about the algorithm's impact across our most simple of protected classes, race and gender, might have detected the disparate effect much earlier and facilitated a correction. And, in some examples, the deployer of the algorithm was reluctant to acknowledge (or flat-out denied) the possibility of bias.[30] Again, this is a problem that is not limited to AI. But as in other instances of unintended bias, we must be able to admit that it might occur, despite our best efforts, and continually monitor decisions to detect it.

Proxy Discrimination

The fourth pernicious flaw that we see at work over and again in recent examples of algorithmic bias is a problem scholars have termed "proxy discrimination."[31] Proxy discrimination occurs when "the predictive power of a facially neutral characteristic is at least partially attributable to its correlation with a suspect classifier."[32] The algorithms identify seemingly neutral characteristics to create groups that closely mirror a protected class, and these "proxies" are used for inclusion or exclusion.

Facebook's use of Lookalike Audiences to facilitate housing discrimination presents one of the clearest illustrations of proxy discrimination. According to allegations by the Department of Housing and Urban Development (HUD), Facebook offered customers that were advertising housing and housing-related services a tool called "Lookalike Audiences." [33] An advertiser using this tool would pick a "Custom Audience" that represented her "best existing customers." Facebook then identified users who shared "common qualities" with those customers, and these similar users become the ad's eligible audience.

To generate a Lookalike Audience, Facebook considered proxies that included a user's "likes," geolocation data, on and offline purchase history, app usage, and page views.[34] Based on

[29] LinkedIn discontinued this practice after some of the reporting. Erica Schwiegershausen, "LinkedIn Will No Longer Ask If You Meant to Search for a Man Instead," *New York: The Cut* (Sept. 8, 2019), https://www.thecut.com/2016/09/linkedin-denies-gender-bias-problem.html.

[30] *Id.*

[31] *See generally* Anya Prince & Daniel B. Schwarcz, "Proxy Discrimination in the Age of Artificial Intelligence and Big Data," *Iowa L. Rev.* (forthcoming), https://ssrn.com/abstract=3347959.

[32] *Id.* at 4-5.

[33] Charge of Discrimination at 4, Facebook, Inc., FHEO No. 01-18-0323-8 (Mar. 28, 2019); *see also* Tracy Jan & Elizabeth Dwoskin, "HUD Is Reviewing Twitter's and Google's Ad Practices As Part of Housing Discrimination Probe," *Washington Post* (Mar. 28, 2019, 6:59 PM), https://www.washingtonpost.com/business/2019/03/28/hud-charges-facebook-with-housing-discrimination/.

[34] *Id.* at 5.

these factors, Facebook's algorithm created groupings that aligned with users' protected classes. Facebook then identified groups that were more or less likely to engage with housing ads and included or excluded them for ad targeting accordingly. According to HUD, "by grouping users who 'like' similar pages (unrelated to housing) and presuming a shared interest or disinterest in housing-related advertisements, Respondent's mechanisms function just like an advertiser who intentionally targets or excludes users based on their protected class."[35]

This is a problem that may persist across advertising algorithms, which are designed to maximize clicks. Even when the advertiser requests a broad audience and more inclusivity, an algorithm will often skew ads to demographic segments that are expected (based on historical performance) to generate more clicks. In one recent study, researchers specified an identical audience for three different job postings: a lumber industry position, a supermarket cashier position, and a taxi position.[36] Despite the request for the same audience, the lumber job went to an audience that was 72% White and 90% male, the supermarket cashier went to an 85% female audience, and the taxi position went to a 75% Black audience.[37]

We have every reason to believe that the dangers of proxy discrimination, amplified by machine learning and optimization, affect the credit sphere as well.[38] The combination of an expanding and innovative FinTech market paired with alternative credit scoring has the potential to extend credit to more folks who need it. But FinTech innovations can also enable the continuation of historical bias, now automated and obscured, to deny access to the credit system and to efficiently target high-interest products to those who can least afford them.[39]

A recent study published by UC–Berkeley scholars illustrates both the promise and residual peril of algorithmic lending decisions.[40] The Berkeley study found that, in loans made by face-to-face lenders, Latinx and African-American borrowers pay considerably more in interest for home-purchase and refinance mortgages as a result of discrimination.[41] The study also found that FinTech algorithms discriminate 40% less—but that significant discrimination

[35] *Id.* at 5-6.

[36] Muhammad Ali et al., "Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes," (arXiv: 1904.02095, 2019), https://arxiv.org/pdf/1904.02095.pdf.

[37] *Id*. A prior study on Google's delivery of job ads demonstrated similar problematic results: in an identical sample that was randomly assigned a male or female identity, Google showed an ad for "$200k+ executive position" to the male group 1852 and just 318 times to the female group. Amit Datta, Michael Carl Tschantz, & Anupam Datta, "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination" (arXiv: 1408.6491 2015), https://arxiv.org/abs/1408.6491.

[38] *See* John Detrixhe & Jeremy B. Merril, "The Fight Against Financial Advertisers Using Facebook for Digital Redlining, *Quartz* (Nov. 1, 2019), https://qz.com/1733345/the-fight-against-discriminatory-financial-ads-on-facebook.

[39] These concerns have also caught Congress's attention. *See, e.g.*, *Examining the Use of Alternative Data in Underwriting and Credit Scoring to Expand Access to Credit*, *Before the H. Comm. On Financial Serv.*, 116th Cong. (2019), https://financialservices.house.gov/calendar/eventsingle.aspx?EventID=404003.

[40] Robert Bartlett et al., "Consumer-Lending Discrimination in the FinTech Era" (Nat'l Bureau of Econ. Research, Working Paper No. 25943, 2019), http://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf.

[41] By 7.9 and 3.6 basis points, respectively. *Id.* at 5.

harming the Latinx and African-American borrowers still occurs.[42] The scholars could not conclude definitively what caused the discriminatory outcomes from the FinTech platforms, but they surmised it was likely due to some type of optimization based on a neutral characteristic that aligned with minority status, just as we saw in the examples above.[43]

Proxy discrimination is not a new problem—the use of facially neutral factors that generate discriminatory results is something our society and our civil rights laws have been grappling with for decades. In the context of AI, sometimes this structural flaw is genuinely accidental. For example, proxy discrimination was one of the reasons that the healthcare algorithm I discussed earlier ultimately produced biased outcomes, but we have no reason to believe that the hospital or manufacturer of the algorithm in question were *trying* to disadvantage Black patients. However, it is important to note that proxy discrimination can also be both intentional and nefarious, because the obscurity provided by black-box decision-making can allow bad-faith actors to effectively launder bias and discrimination through their algorithms in pursuit of illegitimate profits or maintaining oppressive hierarchies.

Thwarting proxy discrimination in the area of AI requires intervention at every step. First, companies must be vigilant about evaluating their inputs, taking special care to avoid proxy-rich datasets in certain contexts. Then they must work tirelessly to interrupt algorithms from using patterns, substitutes, and optimization rates to create systems of "digital redlining" that add only efficiency and opacity to our analog frameworks of economic inequity.

III. Steps toward Algorithmic Justice

There is no question that the four critical algorithmic flaws that we've discussed today—faulty inputs, faulty conclusions, testing failures, and proxy discrimination—have all produced serious harm to consumers and undermine rather than advance economic justice. But I believe these flaws can be prevented or their resulting harms mitigated by smart solutions. Fortunately, many smart thinkers around the globe are grappling with how to develop these effective solutions. As I contemplate steps towards algorithmic justice, the questions I ask are: How does current practice need to change? How can we achieve that change with existing legal and regulatory tools? And what additional or specific tools should be added to our toolbox to tackle these particular problems? In nearly all of the examples I have highlighted, we don't know precisely which inputs and decisions produced the biased or negative outcome. Frustration with the opacity of the "black box" can lead consumers to feel powerless and distrustful.[44] At the same time, the patina of neutral technology making decisions leads to a sense that deployers of bad algorithms are not responsible for the results. The combination of black-box obscurity with the application of complicated and facially neutral technology provides a false sense of security in the objectivity of algorithmic decision-making.

[42] To the tune of 5.3 basis points more in interest for purchase mortgages and 2.0 basis points for refinance mortgages. *Id.* at 6.

[43] In this case, learning that "higher prices could be quoted to profiles of borrowers or geographies associated with low-shopping tendencies." *Id.* at 20.

[44] *See* Jennifer Cannon, "Report Shows Consumers Don't Trust Artificial Intelligence," *FinTech News* (Dec. 4, 2019), https://www.fintechnews.org/report-shows-consumers-dont-trust-artificial-intelligence/.

In other words, while many of the problems of AI—bad data, failure to test, proxy discrimination—have longstanding analogs, AI can simultaneously obscure the problems and amplify them, all while giving the impression that they don't or couldn't possibly exist. Accordingly, the starting point of nearly all discussions about AI ethics and the focal point of many regulatory responses is to require increased transparency and accountability in order to mitigate discriminatory effects.

The European Union's General Data Protection Regulation (GDPR), for example, anchors its AI protections in increased transparency requirements. Under GDPR, the use of automated individual decision-making, including profiling, that produces legal or similarly significant effects triggers certain obligations for data controllers.[45] Controllers must give individuals specific information about the processing, and they must take steps to prevent errors, bias, and discrimination. GDPR also gives individuals the right to challenge and request a review of the decision—sometimes referred to as "the right to an explanation."[46]

This type of transparency is a necessary but not sufficient part of the solution to the pitfalls of AI. It is necessary because it would require the developers and deployers of AI to make sure that AI-generated decisions are explainable and defensible. With the benefit of sunlight, advocates, academics, and other third parties can more widely test for discriminatory and harmful outcomes.[47] And in some cases, more transparency may also empower individuals to challenge incorrect or unfair outcomes themselves.

But transparency alone is not a solution. It must, at a minimum, be coupled with increased accountability and appropriate remedies. Increased accountability means that companies—the same ones who benefit from the advantages and efficiencies of algorithms—must bear the responsibility of (1) conducting regular audits and impact assessments, and (2) facilitating appropriate redress for erroneous or unfair algorithmic decisions. The goal of both transparency and accountability is to limit—or, even better, prohibit—unfair and discriminatory applications of AI.

How best to implement these goals is something that experts and policy leaders are wrestling with at home and internationally, but consensus is building that addressing discrimination is critical to any framework for regulating AI. For example, the EU has issued guidelines that listed seven key requirements that AI systems should meet to be trustworthy,

[45] Regulation 2016/679, art. 22, 2016 O.J. (L 119) (EU).

[46] *Id.* The State of Illinois recently passed a law that seeks to introduce similar transparency into certain hiring decisions. Online Privacy Act of 2019, H.R. 4978, 116th Cong. § 105 (2019) (establishing a right to human review of automated decisions); Artificial Intelligence Video Interview Act, H.B. 2557, 101st Gen. Ass. (Ill. 2019) (enacted) (requiring employers to (1) notify each applicant that AI may be used to analyze the applicant's video interview and consider the applicant's fitness for the position; (2) provide each applicant with information explaining how the AI works and what general types of characteristics it uses to evaluate applicants; and (3) obtain consent from each applicant to be evaluated by the AI program.).

[47] This work is already ongoing but is dependent on the accessibility of data.

including transparency, diversity, non-discrimination and fairness, and accountability.[48]

Here at home, the Office of Management and Budget (OMB) recently issued a memo to executive departments and agencies, entitled "Guidance for Regulation of Artificial Intelligence Applications," setting forth ten principals for agencies to weigh when considering "regulatory and non-regulatory approaches to the design, development, deployment, and operation of AI applications." Again, one of these key principles is "Fairness and Non-Discrimination."[49]

In considering how to apply principles of fairness and non-discrimination, or to achieve goals of transparency and accountability, the first question we must ask is whether and how these ends can be achieved by the application of current law.

IV. Applying Current Law to Better Protect Consumers

Throughout my remarks, I have touched on the theme that the pitfalls of AI-powered algorithms are not wholly different from other problems we have confronted for many years. So our first obligation is to consider how we can apply tried-and-tested solutions to these new fact patterns.

Civil Rights Law

Civil rights laws are the starting point for addressing discriminatory consequences of algorithmic decision-making. Our state and federal civil rights laws already prohibit discrimination in each of the areas we've discussed—healthcare, employment, housing, and credit.[50] None of these laws specifically contemplates discrimination arising in the context of automated decisions relying on vast fields of proxy-rich data. Nor do they allow discrimination simply because it involved an algorithm. "Because AI" is neither an explanation nor an excuse. It is incumbent on law enforcers to think creatively about how to apply existing civil rights law to these new fact patterns; to give credit where it is due, that is exactly what HUD did in its discrimination complaint against Facebook.[51]

[48] *See* High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* 2 (2019), https://ec.europa.eu/futurium/en/ai-alliance-consultation.

[49] *See* Draft Memorandum from the Office of Budget Management to The Heads of Executive Departments and Agencies Concerning Guidance for Regulation of Artificial Intelligence Applications (Jan. 17, 2019), https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf Specifically, OMB advised, "When considering regulations or non-regulatory approaches related to AI applications, agencies should consider, in accordance with law, issues of fairness and non-discrimination with respect to outcomes and decisions produced by the AI application at issue, as well as whether the AI application at issue may reduce levels of unlawful, unfair, or otherwise unintended discrimination as compared to existing processes."

[50] *See, e.g.*, 42 U.S.C. § 2000d et seq. (healthcare); *id.* § 2000e et seq. (employment); *id.* §§ 3601–91 (housing); 15 U.S.C. § 1691 et seq. (credit).

[51] HUD's recent activity in the area of housing discrimination has not all been positive; I harbor substantial concerns about HUD's recent rulemaking activity in this space. Commissioner Chopra recently submitted a comment to HUD with which I agree, that explains the issue and concerns with HUD's approach. *See* Comment of Comm'r Rohit Chopra in the Matter of Proposed Rule to Amend HUD's Interpretation of the Fair Housing Act's Discriminatory

But not all relevant law enforcement agencies have civil-rights authorities—the FTC, for example, has fairly limited explicit enforcement authority in the area of discrimination, yet it is the agency that arguably has the most direct jurisdiction over commercial applications of AI to consumers. And, in many cases, existing civil rights jurisprudence may be difficult to extend to address algorithmic bias precisely because black-box opacity makes establishing discriminatory intent (already a high bar) even more difficult. So we have to consider what other legal protections currently exist.

At my agency, the FTC, there are three types of enforcement authority that provide us with some ability to protect consumers and promote economic justice in the face of algorithmic harms: our general authority under the FTC Act, our sector-specific rules such as FCRA and ECOA, and our rule-making authority under the Magnusson-Moss Act.

Section 5 of the FTC Act

Most of the enforcement conducted by the FTC is brought under the general authority provided to the Commission by Section 5 of the FTC Act, which prohibits unfair or deceptive acts or practices. Our statute is more than a century old, and throughout the agency's history we have been able to apply its general language to meet new enforcement challenges. That same creative thinking urgently needs to be applied to AI.

For example, we could use our deception authority in connection with algorithmic harms where the marketers of algorithm-based products or services represent that they can use the technology in unsubstantiated ways, such as to identify or predict which candidates will be successful or will outperform other candidates. Deception enforcement is well-trod ground for the FTC; anytime a company makes claims about the quality of its products or services, whether or not those products are algorithm-based, the law requires such statements to be supported by verifiable substantiation.

The FTC can also use its unfairness authority to target algorithmic injustice. The unfairness prong of the FTC Act prohibits conduct that causes substantial injury to consumers, where that injury is not reasonably avoidable by consumers and not outweighed by countervailing benefits to consumers or to competition.[52] There are a number of factual predicates that could give rise to an unfairness claim in connection with algorithmic harms. For example, secretly collecting audio or visual data—or any sensitive data—about an individual to feed an algorithm could give rise to an unfairness claim.[53] In addition, if an algorithm is used to

Effects Standard (Oct. 16, 2019), https://www.ftc.gov/system/files/documents/public_statements/1549212/chopra_-_letter_to_hud_on_disparate_impact_proposed_rulemaking_10-16-2019.pdf.

[52] 15 U.S.C. § 45(n).

[53] *See* Fed. Trade Comm'n v. VIZIO, Inc., 2017 U.S. Dist. LEXIS 219381, No. 2:17-cv-00758-SRC-CLW (D.N.J. 2017) (entering judgment); *see also* Press Release, Fed. Trade Comm'n, VIZIO to Pay $2.2 Million to FTC, State of New Jersey to Settle Charges It Collected Viewing Histories on 11 Million Smart Televisions Without Users' Consent (Feb. 6, 2017), https://www.ftc.gov/news-events/press-releases/2017/02/vizio-pay-22-million-ftc-state-new-jersey-settle-charges-it (providing background information).

exclude a consumer from a benefit or an opportunity based on her status in a protected class, such conduct could give rise to an unfairness claim.

I believe that the FTC can and should be aggressive in its use of unfairness to target conduct that harms consumers based on their protected status. But unfairness is an imperfect tool, introducing the hurdles of "reasonable avoidability" and "countervailing benefits" into what can already be a complicated question of the specific injury caused by disparate outcomes.

Vigorous Enforcement of ECOA & FCRA

The FTC also enforces two specific laws that afford protections to consumers related to the extension of credit and their credit information, both of which are relevant to consumers navigating algorithms in the credit sphere. First, the FTC enforces the Equal Credit Opportunity Act (ECOA), which prohibits credit discrimination on the basis of race, color, religion, national origin, sex, marital status, age, or because you receive public assistance.[54] Everyone who participates in the decision to grant credit or in setting the terms of that credit, including real estate brokers and auto dealers who arrange financing, must comply with ECOA. If lenders are using proxies to determine groups of consumers to target for high-interest credit and such proxies overlap with protected classes, the FTC should investigate and, if appropriate, pursue ECOA violations.

A bolder approach would be to incentivize creditors to make use of the ECOA exception that permits the collection of demographic information to test their algorithmic outcomes. Regulation B, which implements ECOA, presumptively prohibits the collection of protected-class demographic information, unlike Regulation C, which implements the Home Mortgage Disclosure Act (HMDA), and generally requires collection of all demographic data for mortgages. The result of these rules is that mortgage credit is monitored closely in a race-conscious way, but all other credit is supposed to go unmonitored. The benevolent idea behind ECOA, of course, was that gender- and race-blind lending would eliminate gender and race disparities. If only that were the case; our longstanding and widespread lived experience shows that gender and race disparities substantially persist, often because of proxy discrimination. I believe that, as with mortgage data, all other kinds of credit should be monitored by creditors consciously for disparities on the basis of protected classes.

Happily, in Regulation B, there is already an exception for collecting demographic data when it is "for the purpose of conducting a self-test,"[55] which is defined as any inquiry "designed and used specifically to determine the extent or effectiveness of a creditor's compliance with the Act or this part."[56] In short, ECOA permits, and the FTC should encourage, non-mortgage creditors to collect demographic data on most borrowers and use it to reduce disparities and train AI and other algorithmic systems to reduce disparities.

[54] 15 U.S.C. § 1691 et seq.

[55] 12 C.F.R. § 1002.5(b)(1).

[56] *Id.* at § 1002.15(b)(1)(i).

Vanishingly few creditors take advantage of this exception.[57] As an enforcer, I will see self-testing as a strong sign of good-faith efforts at legal compliance, and I will see a lack of self-testing as indifference to alarming credit disparities. Of course, if creditors do collect this data to conduct self-testing, they must be able to show that they are not also using it for impermissible purposes.

The FTC also enforces the Fair Credit Reporting Act (FCRA), which applies to companies that compile and sell consumer reports (CRAs) containing consumer information that is used or expected to be used for credit, employment, insurance, housing, or other similar decisions about consumers' eligibility for certain benefits and transactions. FCRA imposes a number of requirements on CRAs to ensure that consumer reports are transparent and accurate and that errors can be corrected.[58] Each of these provisions may enable consumers to seek information about an outcome driven by an algorithmic decision, and I am interested in the FTC exploring how FCRA's rights might lead to increased algorithmic transparency in the credit sphere.

Mag-Moss Rulemaking Initiative

As beneficial as these enforcement initiatives would be, they all target problems after they have occurred rather than prohibiting problematic behavior in advance. But there is one other tool we have to address algorithmic harms on a forward-looking basis: our rulemaking authority under the Magnuson-Moss Act. Unlike many of our sister agencies, the FTC does not have general rulemaking authority under the Administrative Procedure Act (APA), which provides a relatively efficient mechanism for rules to be proposed, commented on by the public, and then finalized after consideration of the comments.[59]

The procedures required to issue a rule under Mag-Moss are substantially more cumbersome than under the APA. It requires the additional steps of a pre-rulemaking advance notice and comment period, a special heads-up to Congress, and public hearings, among other logistical constraints. Historically, the Commission has shied away from extensive Mag-Moss rulemaking as not worth the trouble.

But the threats to consumers arising from data abuse, including those posed by algorithmic harms, are mounting and urgent. I think it is imperative for the FTC to take all action within its authority *right now* to protect consumers in this space. This authority includes our Mag-Moss rulemaking, which, although slow and imperfect, is available today and could

[57] Creditors often find it much easier to never ask about race or gender, or to use (like enforcers generally must for non-HMDA credit) the Bayesian Improved Surname Geocode algorithm to proxy for race, national origin, and gender in datasets of borrowers to self-test for disparities and fair-lending risk.

[58] For example, the FTC brought an action alleging a violation of the FCRA against a company that failed to take reasonable steps to ensure the accuracy of its automated tenant screening system. *See* Press Release, Fed. Trade Comm'n, Texas Company Will Pay $3 million to Settle FTC Charges That it Failed to Meet Accuracy Requirements for its Tenant Screening Reports (Oct. 16, 2018), https://www.ftc.gov/news-events/press-releases/2018/10/texas-company-will-pay-3-million-settle-ftc-charges-it-failed.

[59] In the 1970s, Congress removed the FTC's general ability to issue consumer protection rules under the APA; instead, it gave us the Magnuson-Moss Act.

generate a rule in this area if Congress ultimately fails to act. At the very least, initiating such a rulemaking would significantly advance the public debate through targeted study, thoughtful commentary, and nuanced proposals.

In the area of algorithmic justice, a Mag-Moss rule might be able to affirmatively impose requirements of transparency, accountability, and remedy. A well-drafted rule could do so in a way that takes into account context and relative risk. This is not an easy endeavor, but it is a valuable one to consider.

V. New Rules and Regulations

While we should explore all the authorities we currently have at our disposal, it is also worth considering where there are gaps that can and should be filled with new legislation at the state or federal level.

Several legislative proposals specifically address the types of transparency and accountability requirements I have discussed, but they are perhaps best illustrated by the proposed federal Algorithmic Accountability Act.[60] The proposed bill would impose a number of new requirements on companies using automated decision-making, mandating that they:

- assess their use of automated decision systems, including training data, for impacts on accuracy, fairness, bias, discrimination, privacy and security;
- evaluate how their information systems protect the privacy and security of consumers' personal information; and
- correct any issues they discover during the impact assessments.

The proposed bill also authorizes the FTC to create regulations requiring companies under its jurisdiction to conduct impact assessments of highly sensitive automated-decision systems.[61] The core insight of the proposed bill, through required impact assessments (IAs), is that vigilant testing and iterative improvements are the fair and necessary cost of outsourcing decisions to algorithms. Or, as I would put it, you can't have AI without IA.

In addition, the United States Congress is currently contemplating a federal privacy law. While privacy legislation may not seem directly applicable to the problems we are discussing today, it can in fact play an important role in addressing algorithmic justice—and it is worth noting that the algorithmic-justice requirements imposed in Europe were done as a part of its privacy law, the GDPR. I have been a vocal advocate for a federal privacy law, and I believe that such a bill should incorporate specific protections, including civil rights provisions, to limit the dangers of algorithmic bias and require companies to be proactive in avoiding discriminatory outcomes.

The privacy bill proposed by Senator Cantwell and several colleagues, the Consumer Online Privacy Rights Act, includes a civil rights provision that seeks to accomplish this type of

[60] Algorithmic Accountability Act, H.R. 2231, S. 1108, 116th Cong. (2019).

[61] *Id.* at § 3(b).

broader protection.[62] The bill prohibits the process or transfer of data on the basis of an individual's actual or perceived protected status for the purpose of marketing in a manner that unlawfully discriminates or otherwise makes the opportunity unavailable to the individual or class of individuals.[63] The proposed bill also prohibits the process or transfer of data in a manner that unlawfully segregates, discriminates against, or otherwise makes unavailable the goods, services, or facilities of any place of public accommodations. Throughout our history, we have seen that there is no substitute for strong civil rights laws that outlaw discrimination outright; AI is no exception.

Finally, in the absence of federal action, we have already seen, and I expect us to continue to see, the states—our laboratories of democracy—propose and adopt innovative solutions.

V. Conclusion

I have spent a fair amount of time discussing ways in which we can act—under current or new law—to expose and address some of the challenges posed by AI driven algorithms in order to facilitate its potential to advance justice. But I also think we need to give serious consideration to whether there are applications of AI that pose such serious risk to justice that a ban or a moratorium might be appropriate and necessary. The EU, for example, is currently considering a five-year moratorium on the use of facial recognition technology in public areas.[64]

Considering bans on particular applications of technology is not something we should take lightly. Strong measures like outright prohibitions necessarily involve tradeoffs; we might be sacrificing innovation potential and even some potential improvements in the distribution of justice in order to protect against injustice. That can be the right thing to do.

To return to where I began, however, we should not ban algorithmic applications "because AI." Nor should we allow all applications of the technology unfettered "because AI." We need to consider context- and consequence-specific applications and tailor remedies appropriately. Thoughtful discussions like the ones that have been going on at this conference today are exactly what we need to identify those nuances and move towards justice in this field—both criminal and civil.

[62] This provision is similar to one first included in the Algorithmic Accountability Act, but is part of proposed comprehensive federal privacy legislation. *See* Consumer Online Privacy Protection Rights Act, S. 2968, 116th Cong. (2019).

[63] The Consumer Online Privacy Rights Act protects a wider range of classes than some other civil rights laws. Those protected classes include actual or perceived race, color, ethnicity, religion, national origin, sex, gender, gender identity, sexual orientation, familial status, biometric information, lawful source of income, and disability. Consumer Online Privacy Protection Rights Act, S. 2968, 116th Cong. § 108(a) (2019).

[64] "Facial recognition: EU considers ban of up to five years," *BBC*, Jan. 17, 2020, https://www.bbc.com/news/technology-51148501.