Title

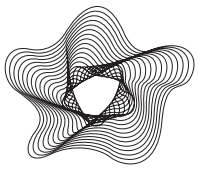
A Survey of Fair and Responsible Machine Learning and Artificial Intelligence: Implications of Consumer Financial Services

Author

Rea, Stephen C

Publication Date

2020-01-28



IMTFI
INSTITUTE FOR MONEY, TECHNOLOGY
& FINANCIAL INCLUSION

A Survey of Fair and Responsible Machine Learning and Artificial Intelligence:

Implications of Consumer Financial Services

By Stephen C. Rea, PhD.
and in partnership with
Capital One's Responsible AI Program

Introduction

Machine learning (ML) algorithms and the artificial intelligence (AI) systems that they enable are powerful technologies that have inspired a lot of excitement, especially within large business and governmental organizations. In an era when increasingly concentrated computing power enables the creation, collection, and storage of "big data," ML algorithms have the capacity to identify non-intuitive correlations in massive datasets, and as such can theoretically be more efficient and effective than humans at using those correlations to make accurate predictions. What is more, AI systems powered by ML algorithms represent a means of removing human prejudices from decision-making processes; since an AI system renders its decisions based solely on the data available, it can avoid the conscious and unconscious biases that often influence human decision-makers.

Contrary to this rosy picture of ML and AI, though, decades of evidence demonstrate how these technologies are not as objective and unbiased as many perhaps wish they were. Biases can be encoded in the datasets on which ML algorithms are trained, arising from poor sampling strategies, incomplete or erroneous information, and the social inequalities that exist in the actual world. And since ML algorithms and AI systems cannot build themselves, the humans who construct them may, however unintentionally, introduce their own biases when deciding on a model's goals, selecting features, identifying which attributes are relevant, and developing classifiers. Additionally, the inherent complexities of ML algorithms that defy explanation even for the most expert practitioners can make it difficult, if not impossible, to identify the root causes of unfair decisions. That same opacity also presents an obstacle for individuals who believe that they have been evaluated unfairly, want to challenge a decision, or try to determine who should—or even *could*—be held accountable for mistakes.

Lingering biases, lack of transparency, and uncertainty regarding accountability fuel public apprehensions about organizations' increasing reliance on ML and AI for making consequential decisions. Such anxieties are understandable, and indeed warranted. But at the same time, misplaced fears could potentially foreclose upon these technologies' beneficial applications. Beliefs about the future of AI can sometimes drift into the realm of science fiction: Will it be like Skynet, the automated defense system that gives way to killer

robots in the *Terminator* franchise? Or will it be a more benevolent technology that automates day-to-day societal operations and provides for the common good? Misconceptions about AI's capabilities derive in large part from confusing it with artificial *general* intelligence (AGI), that is, machines capable of performing intelligence tasks at the level of human cognition that are, at present, purely theoretical. Although contemporary AI systems are much more sophisticated and powerful than when the field of AI research was founded in the mid-1950s, they have not yet reached a point of total autonomy or pure self-awareness. On the one hand, AI's limitations in this respect ought to provide some comfort to those who worry about the prospect of "the machines" rising up to enslave humanity. On the other hand, it means that hopes for AI alone solving the world's social problems and paving the way to a just and equitable future are naïve at best, and dangerous at worst. Managing expectations about what ML and AI can and cannot do is therefore a crucial step in building trust in these systems and making responsible integrations.

This paper surveys current research in and around ML and AI, drawing primarily from work in computer science, social sciences, and the law. Although it examines material across several contexts, the underlying intention is to consider how insights and lessons from a number of different domains can be applied within consumer financial services. And while there are certainly implications for organizational planning and strategy, the analytical focus rests primarily on the individuals and groups who are impacted directly by AI systems' decision-making processes. Compared to other fields, the financial services industry has taken a relatively conservative approach to ML/AI integrations. Consumer-facing applications like robo-advisors for portfolio management, algorithmic trading programs, and proactive marketing tools, as well as harnessing the power of ML to do sentiment analysis of social media feeds and news stories in search of trendlines, have garnered a lot of media attention. So, too, have AI-powered banking assistants like Wells Fargo's Facebook Messenger chatbot and Bank of America's mobile voice assistant "Erica." However, the visibility of initiatives like these in press releases and news items exaggerates their role in financial services today, as they represent less than one-tenth of the funding received in the financial technology, or "fintech," vendor space.[1] Thus far, financial

[1] Daniel Faggella and Raghav Bharadwaj. 2019. *Emerj Vendor Scorecard and Capability Map: AI in Banking 2019*. Boston: Emerj Artificial Intelligence Research.

institutions have primarily invested in ML and AI for automating routine, back-office tasks, improving fraud detection and cybersecurity, and making regulatory compliance easier. Some examples of existing integrations include: JPMorgan Chase's COiN platform, which uses ML for interpreting commercial loan agreements; Citibank's partnership with Feedzai to develop real time risk management through a monitoring platform for payments transactions; and BBVA UK's use of Wolters Kluwer's OneSumX product for regulatory compliance reporting.

The current state of ML and AI in consumer financial services, then, is one in which there is still enormous opportunity for innovation, but also reasons to be cautious. For example, a growing number of fintech lending platforms are using "alternative" data (e.g., utility payment and rental histories, debit card transactions, and auto title and payday loan activities) to develop more inclusive credit scoring algorithms. Alternative credit scoring products are relatively new, but there is already some evidence that they can more accurately assess the risks posed by borrowers with little-to-no credit histories than traditional scoring techniques.[2] However, as law professor Kristin Johnson explains, "Certain algorithms may give occupations like migratory work or low-paying jobs low scores. If a majority of people working in those fields are minorities, the discriminatory result is an unfair impact on those consumers' credit applications."[3] An experimental study of ML in credit risk assessments found that although ML models demonstrated improved accuracy in predicting default rates, they also disproportionately rewarded white borrowers with lower predicted default probabilities and penalized black and Latinx borrowers.[4] Another study of consumer loan origination platforms found that while fintech lenders discriminated less often than face-to-face lenders when it came to accepting or rejecting applicants, they performed about the same with respect to discriminatory pricing of loans: Both face-to-face and online or mobile app-based fintech lenders charged black and Latinx borrowers 6-9 basis points higher on their interest rates than white borrowers, resulting in an extra $250 to $500 million per year in interest payments. The study's authors concluded,

[2] Julapa Jagtiani and Catharine Lemieux. 2018. *The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the LendingClub Consumer Platform*. Federal Reserve Bank of Philadelphia, Working Paper 18-15 (April).
[3] (forthcoming). "Digital Debt." *UC Irvine Law Review* 101-151. P. 150.
[4] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Angsar Walther. 2018. "Predictably Unequal? The Effects of Machine Learning on Credit Markets." (November 6). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3072038.

"With algorithmic credit scoring, the nature of discrimination changes from being primarily concerned with human biases—racism and in-group/out-group bias—to being primarily concerned with illegitimate applications of statistical discrimination."[5] In the words of former Counselor to the Secretary of the Treasury Antonio Weiss, "Just because a credit decision is made by an algorithm, does not mean it's fair."[6]

Weiss's point about fairness applies just as well to consumer financial services beyond credit and lending, as well as ML/AI integrations in other domains. To paraphrase the feminist geographer Doreen Massey, some individuals and groups are more on the "receiving end" of these technologies than others.[7] In other words, ML and AI's advantages and disadvantages are not equally distributed. Nor are the vulnerabilities entailed by digital surveillance techniques for data creation and collection, the sorts of harm that can occur from an erroneous data entry and the burden for correcting it, or the ability to affect how an algorithm interprets one's individual attributes and characteristics. In many ways, ML/AI research's most important contributions have been demonstrating the extent to which structural inequalities—that is, conditions by which one or more groups of people are afforded unequal status and/or opportunities in comparison to other groups—persist by providing quantifiable, documented evidence of social disparities. If an organization's reason for integrating ML- and AI-powered systems is to improve its decision-making procedures so as to make them both more accurate and fairer, then it is imperative to understand and account for persistent inequalities in the social contexts where those systems are embedded. Furthermore, assessing how exactly an algorithmic and/or automated decision-making system could impact specific populations, the risk that it could violate legal standards prohibiting discrimination, and the extent to which the system could perpetuate structural inequalities are of the utmost importance when deciding whether or not to make the integration in the first place.

[5] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. 2018. "Consumer-Lending in the Era of FinTech." NBER Working Paper No. 25943. P. 2.
[6] US Department of the Treasury. 2015. "Remarks by Counselor Antonio Weiss at the Information Management Network Conference on Marketplace Lenders" (October 29). Accessed June 22, 2019: https://www.treasury.gov/press-center/press-releases/Pages/jl0238.aspx.
[7] Doreen Massey. 1994. *Space, Place, and Gender*. Minneapolis: University of Minnesota Press. The original quote reads, "Different social groups have distinct relationships to this anyway differentiated mobility: some people are more in charge of it than others; some initiate flows and movement, others don't; some are more on the receiving end of it than others; some are more effectively imprisoned by it" (149). I am indebted to Steve Jackson for drawing my attention to this passage.

This paper is organized as follows: Section I explores the social contexts with which ML and AI technologies are integrated, and the structural inequalities that influence—and are in turn influenced by—those integrations. Section II surveys ongoing research into data quality, fairness, transparency, and accountability; specific examples of problems that have emerged around these issues; and some of the methods and tools that have been proposed for managing those problems. Finally, the conclusion examines several actual-world cases of ML and AI's human impacts and the challenges and opportunities posed by algorithmic governance. Although in practice these challenges and the ways in which they have been addressed in technological applications can never be totally separated, organizing the paper in this way helps to surface semantic gaps among different stakeholders, and reinforces ML/AI's capacity limitations in their current states.

The examples, issues, and debates discussed in the following sections are by no means exhaustive. And since research in these fields is fast-moving, some are likely to become obsolete while new ones that cannot yet be imagined will almost certainly arise. References in the footnotes provide suggestions for further reading on topics that are of particular interest and not adequately addressed here. With those caveats in mind, this paper offers a starting point for approaching fair and responsible ML/AI integrations in the financial services industry, framing the problems to which they can be addressed, and managing expectations for what they can and cannot do.

Section I: Social Contexts and Structural Inequalities

ML and AI technologies do not operate in a vacuum. No matter the specifics of their design or purpose, algorithms and the systems that they support are always-already embedded in social contexts that affect and are affected by them in equal measure. As Madeleine Clare Elish and Alexandra Mateescu of the Data & Society Research Institute have argued, AI systems are not so much deployed into the world as they are *integrated* with other sociotechnical systems.[8] Understanding ML and AI as integrations rather than deployments highlights the degree to which they are inseparable from broader social and cultural

[8] Alexandra Mateescu and Madeleine Clare Elish. 2019. *AI in Context: The Labor of Integrating New Technologies* (January) New York: Data & Society.

processes, and motivates crucial questions about their broader implications: With what are they being integrated? To what ends? And with what kinds of impact?

Diversity and Representation

To begin with, ML algorithms and AI systems are designed by humans whose biases, however unconscious they may be, are symptomatic of broader structural inequalities affecting representation in development teams. ML research communities have significant diversity problems with respect to both the pipeline and retention of talent.[9] Researchers at the AI Now Institute argue that patterns of exclusion in those communities "[affect] how AI companies work, what products get built, who they are designed to serve, and who benefits from their development."[10] According to Black in AI co-founder Timnit Gebru, without a diversity of perspectives being represented in these fields, "we are not going to address the problems that are faced by the majority of people in the world. When problems don't affect us, we don't think that they're important, and we might not even know what these problems are, because we're not interacting with the people who are experiencing them."[11]

Gebru raises an important point about who gets to be "at the table," so to speak, not only in the expert communities that develop, procure, implement, and regulate ML and AI, but also with respect to those on the receiving end of these technologies. Some ML and AI projects are finding ways to incorporate non-expert voices into the design process. For instance, Desmond U. Patton, Director of Columbia University's SAFElab, writes, "My colleagues in [computer science] and I became keenly aware of what we didn't know early on and knew

[9] For example, only an estimated 12% of the world's leading ML researchers are women (Tom Simonite. 2018. "AI is the Future—But Where are the Women?" *WIRED* (August 17). Accessed June 24, 2019: https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/), and only 18% of the authors whose work was published at the top ML conferences in 2018 were women (JF Gagne. 2019. "Global AI Talent Report." *Jfgagne.ai*. Accessed June 24, 2019: https://jfgagne.ai/talent-2019/). Black and Latinx employees especially are underrepresented in AI development at the largest tech firms: black workers represent 2.5% of Google's full-time workforce, 4% of Facebook's, and 4% of Microsoft's, while Latinx workers account for 3.6% of Google's employees, 5% of Facebook's, and 6% of Microsoft's (Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. *Discriminating Systems: Gender, Race, and Power in AI* (April) New York: AI Now Institute. P. 11).

[10] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. *Discriminating Systems: Gender, Race, and Power in AI* (April) New York: AI Now Institute. P. 5.

[11] Quoted in Jackie Snow. 2018. "'We're in a Diversity Crisis: Cofounder of Black in AI on What's Poisoning Algorithms in our Lives." *MIT Technology Review* (February 14). Accessed June 24, 2019: https://www.technologyreview.com/s/610192/were-in-a-diversity-crisis-black-in-ais-founder-on-whats-poisoning-the-algorithms-in-our/.

that in order to … even begin thinking about what AI could do in this space, we must have community support and buy-in. We not only hire community members and create advisory teams, but privilege their suggestions, critiques and ideas at every turn."[12] Soliciting input from non-experts can help to reveal problems that have not yet been considered by the design and development team, which in turn can work to facilitate trust and understanding among different stakeholder groups. Moreover, interacting with people outside of design labs and boardrooms provides important contextual information about the social dynamics and inequalities that exist in the contexts where AI integrations are made.

To that end, having a complete understanding of locality and history can be vital to the integration's effectiveness, and to being able to strategically account for conflicting perspectives. For example, in late 2017 Boston's public-school system announced that it would be integrating an algorithm designed by MIT researchers to stagger start times at the city's schools in an effort to optimize bus schedules and reduce transportation costs. The announcement was met with protests from parents' groups, and an op-ed in the *Boston Globe* accused the city government of engaging in an undemocratic process of public service design.[13] In light of the public outcry, Boston Public Schools scrapped the plan. However, the *Globe* op-ed's authors later met with the MIT research team and discovered that not only were the groups protesting the changes unrepresentative of the affected population—the vast majority of the plan's opponents were affluent, white families, who accounted for only about 15% of Boston's public school students—but also that the algorithm's developers had engaged in substantive community outreach and designed it specifically to help lessen the existing scheduling system's burden on low-income families. These parts of the design process had not been communicated in the Boston Public Schools' announcement, and so there was limited public understanding of the logics that motivated the integration.[14] The Boston case offers an important lesson about the contextual nuances of structural inequality, and how even if diverse perspectives are

[12] Desmond U. Patton. 2019. "Why AI Needs Social Workers and 'Non-Tech' Folks." *Noteworthy - The Journal Blog* (March 24). Accessed June 25, 2019: https://blog.usejournal.com/why-ai-needs-social-workers-and-non-tech-folks-2b04ec458481, emphasis removed.
[13] Kade Crockford and Joi Ito. 2017. "Don't Blame the Algorithm for Doing What Boston School Officials Asked." *The Boston Globe* (December 22). Accessed August 9, 2019: https://www3.bostonglobe.com/opinion/2017/12/22/don-blame-algorithm-for-doing-what-boston-school-officials-asked/lAsWv1Rfwqmq6Jfm5ypLmJ/story.html?arc404=true.
[14] Joi Ito. 2018. "What the Boston School Bus Schedule Can Teach Us about AI." *WIRED* (November 5). Accessed June 25, 2019: https://www.wired.com/story/joi-ito-ai-and-bus-routes/.

sought out, this alone is not enough to build stakeholder trust without transparency and effective communication.

The same structural inequalities that are reflected in research fields and design processes also pertain to how ML/AI technologies operate. Unlike their more deterministic predecessors, one of ML algorithms' greatest advantages is that they have the capacity to learn from multiple iterations and refine themselves over time. Whether supervised or unsupervised, ML processes need training data to get started. Training data, by necessity, are drawn from historical datasets (although, many algorithms can now also learn from real time data monitoring). In many ways, those datasets are records of structural inequality: It is through data that ML techniques "can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society."[15] For example, the legacy of racism embedded in the United States' social hierarchies has ripple effects that can be observed in data about criminal justice sentencing, employment and hiring patterns, and housing and credit discrimination, to name but a few.[16]

The Racial Wealth Gap

In the context of consumer financial services, arguably the most significant structural inequality is the racial wealth gap that exists between black and white Americans. This gap is observable at the lowest and highest ends of the economic spectrum: white households living near the poverty line have, on average, around $18,000 in wealth, while black households in the same position have median wealth near $0; and black households make up less than 2% of the top 1% of America's wealth distribution, compared to the more than

[15] Solon Barocas and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104 (3): 671-732. P. 674.

[16] It also bears acknowledging that race as a concept and as identity is purely a social construction, the product of specific histories of political organization. Computer scientist Sebastian Benthall and sociologist Bruce D. Haynes argue, "Because race is not an inherent property of a person but a 'social fact' about their political place and social location in society, racial statistics do not reflect a stable 'ground truth'. Moreover, racial statistics by their very nature mark a status inequality, a way of sorting people's life chances, and so are by necessity correlated with social outcomes" (2019. "Racial Categories in Machine Learning." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 289-298. P. 290). Since race is a significant vector for structural inequality that is reflected in training data and can be reified through algorithmic outputs, understanding the specific ways in which race functions as a means of social organization and discrimination is crucial to assessing its use and misuse in ML.

96% of wealthiest Americans who come from white households.[17] The racial wealth gap is preserved in part through discrimination in credit and lending. Economic growth in the immediate post-World War II United States was driven in large part by government-subsidized credit markets. Faith in credit as a vehicle for socioeconomic mobility has persisted in American policymaking ever since (events such as the 2008 global financial crisis notwithstanding). However, this reading of history belies how the benefits of that growth and access to affordable credit were not equally distributed. Beginning in the New Deal era, white Americans benefited from access to cheap credit and preferential home loans, while black Americans were largely excluded from the political and economic mechanisms that made fair, affordable credit available—either through outright discriminatory laws, or through more indirect methods like redlining—and so had to make do with the predatory fringes of consumer lending.[18] This pattern of disparity demonstrates a crucial aspect of structural inequality: it is not simply a matter of unequal disadvantage, but also of unequal *advantage.* In other words, the benefits that white Americans experienced were produced by the very same legal and economic frameworks that penalized black Americans. As data and information scientist Anna Lauren Hoffman argues, "Instead of treating as morally abhorrent those structural processes that unjustly advantage certain groups, the focus on disadvantage forces us into a kind of benevolent—or, worse, patronizing—stance that flattens our understanding of those already relegated to the 'basement' of the social hierarchy."[19]

Nowhere perhaps did race-based structural inequalities in credit markets sustain the racial wealth gap more than in home ownership, a traditional means of wealth accumulation for American consumers. Denied access to government-backed, low-interest mortgages, prospective black homebuyers were pushed into situations where they had to pay much higher interest rates and were not granted the titles to their homes until their debts had been paid in full. Housing discrimination was also one of the conditions of possibility for de facto neighborhood segregation, which in turn provided the basis for banks' redlining

[17] William Darity Jr., Darrick Hamilton, Mark Paul, Alan Aja, Anne Price, Antonio Moore, and Caterina Chiopris. 2018. *What We Get Wrong About Closing the Racial Wealth Gap*. Samuel DuBois Cook Center on Social Equity and Insight Center for Community Economic Development (April). P. 2.
[18] Abbye Atkinson. 2019. "Rethinking Credit as Social Provision." *Stanford Law Review* 71 (5): 1093-1162.
[19] 2019. "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse." *Information, Communication & Society* 22 (7): 900-915. P. 906.

practices.[20] Law professor Mehrsa Baradaran argues that these conditions enabled a self-reinforcing debt cycle:

> "Over 70 percent of suburban black families had to borrow just so they could purchase cars, appliances, furniture, and other life necessities. Because the black middle class had more debt, they were charged higher interest on each new loan. More debt begets higher interest and vice versa. The added debt burden and high interest was a direct result of the lack of wealth, and, looping around once again, the debt made it even harder to accumulate more wealth. The debt-wealth cycle fed on itself. Black middle-class families making the same incomes as the white middle class had much less wealth—a disparity that both created their need for debt and was caused by the costly debt."[21]

As wealth, debt, and property are often transferred generationally, these patterns of discrimination persist and affect borrowers' opportunities in the present day. For example, researchers at the National Fair Housing Alliance (NFHA) have documented the long-term effects of the "dual credit market" that grew out of housing segregation and redlining, noting that in the early 2000s black and Latinx borrowers received high-interest subprime loans at a rate nearly three times that for white borrowers; what is more, non-white borrowers were more likely to be offered subprime loans even if their credit scores qualified them for low-interest loans.[22] Subprime loans' high interest rates make repayment

[20] See Richard Rothstein. 2017. *The Color of Law: A Forgotten History of How Our Government Segregated America*. New York: Liveright.
[21] 2017. *The Color of Money: Black Banks and the Racial Wealth Gap*. Cambridge, MA: The Belknap Press of Harvard University Press. Pp. 112.
[22] Lisa Rice and Deidre Swesnik. 2012. *Discriminatory Effects of Credit Scoring on Communities of Color*. Prepared for the Symposium on Credit Scoring and Credit Reporting Sponsored by Suffolk University Law School and National Consumer Law Center (June 6-7). Two high-profile cases in the wake of the 2008 global financial crisis exemplify this type of discrimination. Between 2004 and 2009, both Countrywide Financial Corporation and Wells Fargo pushed black and Latinx borrowers who had qualified for prime mortgage loans off onto subprime loans and loans with higher fees and rates than for similarly qualified white borrowers. Both organizations settled with the Department of Justice for $335 million and $184.3 million respectively (US Department of Justice. 2011. "Justice Department Reaches $335 Million Settlement to Resolve Allegations of Lending Discrimination by Countrywide Financial Corporation" (December 21). Accessed July 10, 2019: https://www.justice.gov/opa/pr/justice-department-reaches-335-million-settlement-resolve-allegations-lending-discrimination; 2012. "Justice Department Reaches Settlement with Wells Fargo Resulting in More Than $175 Million in Relief for Homeowners to Resolve Fair Lending Claims" (July 12).

more difficult. Failure to make monthly payments on a loan can negatively impact a credit score, which in turn means that those borrowers are ineligible for lower interest rate loans, trapping them in debt cycles from which it can be difficult to escape. In other words, "a borrower may well end up with a damaged credit score not because the borrower was more risky or negligent but rather because the borrower obtained a loan through a broker or received loan terms that increase the likelihood of delinquency and default."[23] Oftentimes the most readily available options for subprime borrowers are payday lenders that charge interest rates as high as 667%,[24] and tend to be more concentrated on average in neighborhoods of color.[25] Histories of racial discrimination in credit and neighborhood segregation are thus encoded in the data that are used for training ML algorithms for evaluating credit risk. Without some way of accounting for those structural inequalities and mitigating their effects, algorithmic assessments simply reproduce the same disparities in their outcomes.

Policing, Hiring, Immigration

Consumer finance is not the only domain in which structural inequalities affect ML algorithms' training data, and by extension AI systems' decisions and outcomes. For example, PredPol, a predictive policing algorithm that is used by Los Angeles, Seattle, and Santa Cruz, among other cities, uses quantitative statistical analysis to forecast areas where property crimes are likely to occur. Municipal law enforcement agencies can use these predictions to allocate resources and officers in ways that, theoretically, are more objective than traditional policing practices. However, the data that PredPol uses to make its

Accessed July 10, 2019: https://www.justice.gov/opa/pr/justice-department-reaches-settlement-wells-fargo-resulting-more-175-million-relief.

[23] Lisa Rice and Deidre Swesnik. 2012. *Discriminatory Effects of Credit Scoring on Communities of Color*. Prepared for the Symposium on Credit Scoring and Credit Reporting Sponsored by Suffolk University Law School and National Consumer Law Center (June 6-7). P. 18.

[24] Leonhardt, Megan. 2018. "This Map Shows The States Where Payday Loans Charge Nearly 700 Percent Interest." *CNBC.com* (August 3). Accessed June 27, 2019: https://www.cnbc.com/2018/08/03/states-with-the-highest-payday-loan-rates.html.

[25] Wei Li, Leslie Parrish, Keith Ernst, and Delvin Davis. 2009. *Predatory Profiling: The Role of Race and Ethnicity in the Location of Payday Lenders in California*. Center for Responsible Lending (March 26). In theory, payday loans are tools for managing financial shocks. However, empirical evidence shows that most payday loan borrowers use them to pay off basic living expenses such as rent, food, and utilities (Abbye Atkinson. 2019. "Rethinking Credit as Social Provision." *Stanford Law Review* 71 (5): 1093-1162; see also Lisa Servon. 2017. *The Unbanking of America: How the New Middle Class Survives*. New York: Houghton Mifflin Harcourt).

predictions come from historical records that may include incorrect labels, sampling bias, and skewed feature selection due to policing practices that overestimate the criminality of non-white communities.[26] Some have attributed this unevenness to "the culture of data production in policing" that produces "data that is derived from or influenced by corrupt, biased, and unlawful practices, including data that has been intentionally manipulated or 'juked,' as well as data that is distorted by individual and societal biases."[27] As such, "it is clear that police records do not measure crime. They measure some complex interaction between criminality, policing strategy, and community-police relations."[28] By relying on biased datasets, then, predictive policing algorithms like PredPol contribute to the maintenance of race-based structural inequalities.

Predictive algorithms are also increasingly used in hiring processes as a means of more efficiently screening applicants' resumes and targeting job advertisements at desirable candidates. Additionally, many recruiters approach the use of algorithmic hiring tools as a way to remove interpersonal biases and thereby more effectively match job candidates with positions to which they are well suited. However, hiring algorithms can perpetuate the same structural inequalities as predictive policing and credit risk assessment tools when they are trained on data that reflect workplace biases and preferences for certain types of individuals over others.[29] A recent example of this was Amazon's decision to do away with its AI-based recruiting tool after it showed a bias for male applicants in software development and other tech-heavy roles. Amazon's ML specialists determined that the algorithm, which had been trained on a decade's worth of résumé submissions, had learned to downgrade female applicants based on natural language processing that favored "masculine-sounding" words and phrases.[30]

Algorithms are also being used in US immigration vetting protocols for analyzing biometric, biographic, and social media identification data. Law professor Margaret Hu has

[26] Andrew D. Selbst. 2017. "Disparate Impact in Big Data Policing." *Georgia Law Review* 52 (1): 109-195.
[27] Rashida Richardson, Jason M. Schultz, and Kate Crawford (forthcoming). "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice." *New York University Law Review.* P. 195.
[28] Kristian Lum and William Isaac. 2017. "To Predict and Serve?" *Significance* 13 (5): 14-19. P. 16.
[29] Miranda Bogen and Aaron Rieke. 2018. *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias* (December). Washington, D.C.: Upturn.
[30] Jeffrey Dastin. 2018. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." *Reuters* (October 9). Accessed June 27, 2019: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

characterized these efforts as tantamount to an "algorithmic Jim Crow" regime, referencing one of the most notorious legal justifications for structural inequality in American history. She argues that under algorithmic Jim Crow, "rather than relying upon a targeted class, such as race, national origin, gender, or religion, as a sole basis for exclusion, big data allows for exclusion to be based on an abstraction, such as digitally inferred or algorithmically anchored guilt or suspicion."[31] Whereas credit risk assessments, predictive policing, and hiring algorithms demonstrate how biased data and inaccurate labeling can reproduce structural inequalities, the situation that Hu describes illustrates the importance of attending to the sociopolitical contexts with which AI integrations are made and how algorithms can be incorporated into legal mechanisms for discrimination. As she elaborates, the separation enabled under algorithmic Jim Crow "is achieved through data discrimination applied on the back end of screening and vetting protocols rather than overt social and economic discrimination and legal apartheid on the front end of segregationist regimes."[32] This is especially troubling because it means that a process that on its face complies with legal standards mandating fair and equal treatment can still produce outcomes aligned with prejudiced organizational goals, thereby demonstrating the limitations of regulation for combating discrimination through the use of ML/AI technologies.

Where Do We Go From Here?

Some researchers and policymakers hold out hope that, with continued improvements to ML and AI's accuracy and fairness, automated technologies will help to facilitate a more equitable future. Others, however, are less optimistic. As law and technology experts Julia Powles and Helen Nissenbaum note, "Bias is a social problem, and seeking to solve it within the logic of automation is always going to be inadequate."[33] They assert that focusing solely on building more effective, fairer AI is ultimately a distraction from addressing the social foundations of structural inequality. Others take this criticism further, emphasizing how AI integrations are not only incapable of fixing fundamentally broken institutional processes,

[31] 2017. "Algorithmic Jim Crow." *Fordham Law Review* 86 (2): 633-696. P. 658.
[32] *Ibid.* P. 695.
[33] 2018. "The Seductive Diversion of 'Solving' Bias in Artificial Intelligence." *Medium* (December 7). Accessed June 27, 2019: https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53.

but that they can, even inadvertently, exacerbate the structural inequalities that those processes uphold.[34] In this vein, artist and researcher Mimi Onouha has coined the term "algorithmic violence" to refer to "violence that an algorithm or automated decision-making system inflicts by preventing people from meeting their basic needs"; she elaborates, "[Forms of algorithmic violence] not only affect the ways and degrees to which people are able to live their everyday lives, but in the words of Mary K. Anglin, they 'impose categories of difference that legitimate hierarchy and inequality.'"[35] Critics in this camp also advocate rethinking the data collection and creation practices that make ML and AI possible by recognizing the inequalities involved in decisions about what counts as data, whose data matter, and how those data are put to use, and attending to the unequal distribution of vulnerability and harm with respect to automation's effects. For example, while AI-powered facial recognition software's shortcomings—such as the now infamous example of Google Photos' labeling of black faces as "gorillas"[36]—have prompted efforts to improve these systems' accuracy and train them on more inclusive datasets, critics point to how the use cases for these technologies include police surveillance practices, which have historically been targeted at the poor and people of color.[37] Attending to the power dynamics surrounding these technological integrations, from data creation and collection all the way

[34] To wit, regarding predictive policing algorithms, the Stop LAPD Spying Coalition argues, "The collection of data, of any type, can never escape bias. The collection of data carries an inherent purpose and intention. Historically and currently there exists an intention and purpose of categorizing and documenting acts as criminal. That is, crime is created and enacted into law by those in power in order to serve the interests of the powerful, and as a result, crime data is a reflection of law enforcement's responses to particular kinds of behaviors committed by certain subsets of the population" (2018. *Before the Bullet Hits the Body: Dismantling Predictive Policing in Los Angeles* (May 8). Pp. 13-14).

[35] 2018. "Notes on Algorithmic Violence." *GitHub.com* (February 8). Accessed June 28, 2019: https://github.com/MimiOnuoha/On-Algorithmic-Violence.

[36] Conor Dougherty. 2015. "Google Photos Mistakenly Labels Black People 'Gorillas'." *The New York Times Bits Blog* (July 1). Accessed June 28, 2019: https://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas/?mtrref=undefined.

[37] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. *Discriminating Systems: Gender, Race, and Power in AI* (April) New York: AI Now Institute; see also Sigal Samuel. 2019. "Some AI Just Shouldn't Exist." *Vox* (April 19). Accessed June 28, 2019: https://www.vox.com/future-perfect/2019/4/19/18412674/ai-bias-facial-recognition-black-gay-transgender. China's social credit system is perhaps the most ambitious—and troubling—example yet of how surveillance and AI technologies can be incorporated into governance strategies that distribute advantage and disadvantage unequally, "result[ing] in material benefits and reputational praise or material exclusion and reputational loss" (Severin Engelmann, Mo Chen, Felix Fischer, Ching-yu Kao, and Jens Grossklags. 2019. "Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines 'Good' and 'Bad' Behavior." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 69-78. P. 70).

through to automated decisions and their effects, helps to illustrate how it is that some are more on the receiving end of ML and AI than others.

All of the examples in this section provide evidence for why it is important to examine the social contexts with which ML/AI integrations are made, and to critically analyze their relationships with structural inequalities. They also demonstrate why managing expectations for what AI can and cannot do is a vital part of not only building trust among different stakeholder groups, but also assessing whether or not an integration can be made responsibly and the specific ways in which it could be harmful to individuals and groups. The next section discusses ML and AI's technical affordances and limitations, and how researchers and practitioners are approaching issues related to data, fairness, discrimination, and explainability.

Section II: Technical Perspectives

ML researchers are continually working to improve algorithms' accuracy and fairness, and, where applicable, the decision-making systems that they support. Generally speaking, ML algorithms learn how to map from a set of inputs to some desired output; when they are part of a decision-making process, such as approving a loan, that output is often a binary "yes/no" determination. Arriving at these kinds of decisions—or "classifications"—involves predictive modeling, that is, calculating the probability that something will occur, such as a borrower defaulting on their loan. Since classifications should be as accurate as possible in order to be useful, the model making those predictions needs to be given certain parameters that inform its decision-making, and then train itself on actual-world data to continually refine those parameters.

At the same time, when a decision-making system is integrated with a high-stakes social context, there is a certain expectation that the predictions it makes will be "fair," that is, that they will not be distributed in such a way that they consistently penalize certain individuals and groups while benefiting others. This is the central dilemma for ML: how to map inputs to outputs with a high degree of accuracy, but without also producing discriminatory classifications. If the structural inequalities outlined in the previous section did not exist—that is, if the social contexts within which ML/AI operate afforded an equal playing field for everyone—then striking a balance between accuracy and fairness would be a relatively

simple proposition. Since this is not the case, however, understanding AI systems' technical affordances and limitations is crucial for evaluating the tradeoffs involved in specific contexts of application and being able to perform impact assessments ahead of time regarding their potential benefits and harms. This section explores some of the most pressing technical questions and concerns surrounding ML/AI integrations, including: data quality; fairness and accuracy; mitigating discrimination; model explainability; and recourse mechanisms.

Data

As mentioned in the introduction, one of the conditions of possibility for current investments in ML/AI is the existence of "big data," made possible through the proliferation of surveillance, collection, and recording instruments as well as how, through interaction, humans and technical systems coproduce new forms of data, e.g., search histories, digital transactions, geolocations, etc.[38] As advances in computational processing speeds and information storage infrastructures make it possible to analyze big data more quickly and cheaply, organizations compete with one another to turn analytics into actionable insights. The *quality* of the data being analyzed is critical to being able to produce accurate predictions and avoid erroneous classifications with potentially harmful outcomes. ML algorithms can inherit biases encoded in training data, and incomplete or skewed datasets can negatively affect the accuracy of outputs. But at an even more fundamental level, the fact that ML algorithms can identify non-intuitive correlations in big data can easily lead to "apophenia: seeing patterns where none actually exist, simply because massive quantities of data can offer connections that radiate in all directions."[39] In other words, blind faith in ML's potential to discover relationships that no human analyst could detect can end up overestimating the significance of those relationships. For instance, in the field of sentiment analysis where data such as social media posts and comment sections are analyzed for indicators of emotion and mood, researchers have observed a tendency toward prosopoeia, that is, "attributing an imagined and unified voice to a dispersed and invisible

[38] Bill Maurer. 2015. "Principles of Descent and Alliance for Big Data." In *Data, Now Bigger and Better!*, edited by Tom Boellstorff and Bill Maurer. Pp. 67-86. Chicago: Prickly Paradigm Press.
[39] Danah boyd and Kate Crawford. 2011. "Six Provocations for Big Data." In *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, September 21-24, Oxford, UK. P. 2.

aggregate."[40] Put another way, "If the relationships between different variables correlate in the aggregate, there is danger in assuming the same relationship will also correlate at an individual level—an error known as the 'ecological fallacy.'"[41] Certainly, useful insights can be drawn from aggregated data and used to predict individual behaviors. But if those relationships are assumed to be absolute and used to render decisions in high-stakes social contexts, then falling victim to apophenia, prosopopoeia, and/or ecological fallacies can have destructive consequences for individuals that are difficult, if not impossible, to remedy.[42] While big data analytics afford incredible value for the organizations that can take advantage of them, being able to align the "right" data with organizational goals is crucial for any ML/AI integration. Doing so responsibly entails having a clear understanding of data quality—including what they indicate and what they do not—and the degree to which patterns in aggregate data can or should be applicable in individual cases.

Social media data are especially attractive because of their potential to reveal information that other data sources might fail to capture. However, there are also reasons to beware of mining social media profiles both with respect to accuracy and possible discrimination (not to mention privacy). For example, it has been suggested that social media posts might provide insight into individual behaviors that a lender could use for making credit risk assessments, especially for thin-file borrowers about whom credit rating agencies lack more traditional data. However, as mathematician Cathy O'Neil notes, "Most of us are Facebook friends with a bunch of people from high school. If I went to high school in a poor community but now have the means to pay back my loans, this method could wrongly rule me out."[43] Put another way, "An algorithm that assumes financially responsible people socialize with other financially responsible people may incorporate systemic biases, and

[40] Mark Andrejevic. 2011. "The Work that Affective Economics Does." *Cultural Studies* 25 (4-5): 604-620. P. 612.
[41] Luke Stark. 2018. "Algorithmic Psychometrics and the Scalable Subject." *Social Studies of Science* 48 (2): 204-231. P. 215.
[42] Ping An Puhui, China's second-largest life insurance company, has developed a tool that it claims can use facial recognition analysis to determine the probability that someone will default on a loan, offering a contemporary example of how big data, AI, and automated decision-making might be put toward specious ends (Glen Gilmore. 2017. "Facial Recognition AI Will Use Your Facial Expressions to Judge Creditworthiness." *Medium* (October 29). Accessed July 4, 2019: https://medium.com/@glengilmore/facial-recognition-ai-will-use-your-facial-expressions-to-judge-creditworthiness-b0e9a9ac4174).
[43] 2015. "How to Talk About Big Data and Lending Discrimination." *American Banker*. (September 10). Accessed July 2, 2019: https://www.americanbanker.com/opinion/how-to-talk-about-big-data-and-lending-discrimination.

deny loans to individuals who are themselves creditworthy but lack creditworthy connections."[44] Aracely Panameño, director of Latino Affairs for the Center for Responsible Lending, cautions, "Alternative data is not created equal … It can result in disparate impact, potential racial discrimination, [and] red-lining—meaning that [consumers] might end up being charged more for certain products and services on the basis of where they live."[45] The Fair Isaac Corporation (FICO), developers of the most widely used credit scoring tools in the financial services industry, has warned that "not all [alternative data] provide equal value for scoring,"[46] explaining that in order to be useful—let alone *usable*—those data must capture many variables, be statistically representative and consistent over time, be verifiable, and comply with consumer privacy and antidiscrimination laws.

Fairness & Bias

Together with data quality, fairness constraints for ML must also be aligned with organizational goals. As the AI Now Institute's Director of Policy Research Rashida Richardson argues, it is important to acknowledge the semantic differences that "fairness" has inside and outside of ML communities, and the ways in which those differences have been used to abstract from and oversimplify social and historical contexts like the ones discussed in the previous section.[47] Moreover, understanding semantic gaps with respect to how fairness is used in different contexts is helpful for assessing regulations that govern ML/AI integrations. Work on fairness in ML is motivated primarily by a desire to mitigate the effects of bias—from sampling issues, feature selection, labeling, etc.—and to prevent discrimination in a given model's outputs.[48] In this context, fairness typically denotes one of

[44] Kevin Petrasic, Benjamin Saul, James Greig, Matthew Bornfreund, and Katherine Lamberth. 2017. *Algorithms and Bias: What Lenders Need to Know*. Washington, D.C.: White & Case LLP. P. 4. Other fintech companies have sought to use personal health tracking data, such as data from FitBit or Nike's FuelBand, to better assess consumers' credit risks (Philipp Kallerhoff. 2013. *Big Data and Credit Unions: Machine Learning in Member Transactions*. Madison, WI: Filene Research Institute). As with social media data, the relationship between health data and creditworthiness rests on rather specious assumptions.

[45] Quoted in Colin Wilhelm. 2018. "Big Data vs. the Credit Gap." *Politico* (February 7). Accessed July 1, 2019: https://www.politico.com/agenda/story/2018/02/07/big-data-credit-gap-000630.

[46] 2015. *Can Alternative Data Expand Credit Access?* FICO Decisions Insights White Paper No. 90. P. 7.

[47] This summary of Richardson's argument is based on a presentation given at a closed-door symposium in May 2019, and used here with her permission.

[48] "Discrimination," like fairness, also has semantic differences inside and outside of ML. As information scientist Solon Barocas and attorney Andrew Selbst point out, "By definition, data mining is *always* a form of statistical (and therefore seemingly rational) discrimination. Indeed, the very point of data mining is to provide a rational basis upon which to distinguish between individuals

two things: individual fairness and group fairness. In a particularly influential ML research paper, Cynthia Dwork et al. define individual fairness as the principle by which "any two individuals who are similar with respect to a particular task should be classified similarly."[49] In order to be fair, classifiers should produce similar probabilities for individuals who are relatively "close" to one another with respect to the model's relevant metrics. Group fairness, on the other hand, can be achieved by equalizing a classifier's relevant statistics across every group that the model observes. There are three prominent conditions for group fairness in ML: statistical parity, which requires classification rates to be equal across all groups (e.g., men and women are approved and rejected for loans at equal rates); equalized odds, which requires true positive and false positive rates to be equal across all groups (e.g., men and women are correctly or erroneously approved for loans at equal rates); and calibration, which requires that for any predicted score, the proportion of actual positives is equal across all groups (e.g., men and women receive the same score default at equal rates).

In legal contexts, however, fairness relates directly to antidiscrimination statutes. The US's tradition of jurisprudence has tended to view fairness as a core principle in the goal of society-wide equilibrium of rights, opportunities, and resources. Title VII of the Civil Rights Act of 1964[50] exemplifies this approach by codifying the concepts of "disparate treatment" and "disparate impact" in relation to "protected classes," that is, certain sensitive attributes like race, gender, age, etc. that are commonly used as criteria for differentiating groups from one another. Disparate treatment occurs when individuals and/or groups are subjected to unequal processes on the basis of their protected class status. Prohibitions against disparate treatment, then, are designed to ensure procedural fairness and equal opportunity. Disparate impact arises when unequal outcomes follow discriminatory

and to reliably confer to the individual the qualities possessed by those who seem statistically similar" (2016. "Big Data's Disparate Impact." *California Law Review* 104 (3): 671-732. P. 677). In civil rights discourse, however, it refers to the unequal treatment of individuals and/or groups on the basis of certain characteristics and is the object of legal prohibitions. Statistical discrimination can, of course, beget illegal discrimination, but understanding how they differ is crucial to assessing ML and AI from a technical perspective.

[49] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2011. "Fairness Through Awareness." In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, January 8, Cambridge, MA. Pp. 214-226. P. 215, emphasis removed.

[50] While Title VII explicitly mandates rules for employment and labor practices in the US, the principles of disparate treatment and disparate impact have more far-reaching implications for civil rights law and, by extension, ML and AI.

patterns that are based on protected attributes, even if the process for arriving at those outcomes is, on its face, neutral. Outlawing disparate impact is a means of enacting distributive justice, or the equitable allocation of costs and rewards.[51] Except under special circumstances such as affirmative action or business necessity, decision-making processes must not violate either of these two principles. Legal definitions of fairness impact ML/AI by placing constraints on what algorithms can and cannot do. Specifically, an AI-enabled decision-making system perpetrates disparate treatment if its decisions are based—in whole or in part—on protected attributes, and it has disparate impact if those decisions penalize and/or reward individuals and groups according to patterns that correlate with protected class status, even if the system was not intentionally designed to discriminate in this way. With these constraints in place, an automated decision-maker cannot, in theory, make use of an intentionally discriminatory tool that inflicts harm on some individuals and groups and unfairly benefits others.

Complicating matters further is the fact that ML communities themselves do not have a consensus definition of fairness. This lack of agreement is somewhat beneficial because of the flexibility that it affords programmers; since ML/AI integrations are context- and domain-specific, what is fair in one situation may be unfair in another. But it also underscores how optimizing for fairness depends largely on the goals and desired outcomes for specific use cases. Satisfying different fairness conditions almost always entails some degree of trade-off with respect to accuracy (given that algorithmic decision-makers in the financial services industry work in contexts of structural inequality, it makes sense that the most "accurate" predictions will be those that reflect the fundamental unfairness of those contexts). There are also trade-offs involved in trying to balance the conditions for individual versus group fairness summarized above, each of which has its own shortcomings: Individual fairness suffers from an assumption that there is a universally agreed-upon metric to measure similarity with respect to the task, while group

[51] Some critics have drawn attention to the shortcomings of these legal standards for fairness within actual-world social contexts. Media and technology scholars danah boyd, Karen Levy, and Alice Marwick have argued that anchoring concepts of fairness to a narrow consideration of individuals ignores how we are all ineluctably connected to broader social networks, and so in the context of AI systems, "algorithms that identify our networks, or predict our behavior based on them, pose new possibilities for discrimination and inequitable treatment" (2014. "The Network Nature of Algorithmic Discrimination." In *Data and Discrimination: Collected Essays*, edited by Seeta Peña Gangadharan with Virginia Eubanks and Solon Barocas. Pp. 53-57. Washington, D.C.: Open Technology Institute/New America. P. 56).

fairness relies on averages that may not faithfully reflect the model's behavior near the decision threshold. Researchers have demonstrated that, except in rare cases, it is impossible to satisfy every fairness condition simultaneously,[52] and so it is incumbent upon organizations procuring and integrating algorithmic decision-making systems to clearly identify the sorts of harmful outcomes they wish to avoid. Moreover, optimizing for group fairness can encounter problems when a classifier that performs well for pre-defined groups is unfair to subgroups that were not designated as protected, a circumstance known as "fairness gerrymandering."[53]

From a more philosophical perspective, different fairness definitions amount to "different interpretations of the extent to which factors outside of an individual's control should be factored into decisions made about them and the extent to which abilities are innate and measurable."[54] Whereas US legal traditions regarding fairness are oriented toward ensuring equal treatment and equality of opportunity, abiding by these principles in the context of ML can actually work to *undermine* these principles. Somewhat counterintuitively, researchers have demonstrated how allowing ML algorithms to "[use] sensitive attributes may increase accuracy for all groups and may avoid biases where a classifier favors members of a minority group that meet criteria optimized for a majority group."[55] Although "fairness through unawareness," that is, ignoring sensitive attributes in the interest of

[52] See Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *arXiv preprint arXiv:1609.05807*; and Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. "On the (Im)possibility of Fairness." *arXiv preprint arXiv:1609.07236*.

[53] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness." *arXiv preprint arXiv:1711.05144.* Fairness gerrymandering relates indirectly to the social phenomenon of "intersectionality" (Kimberlé Crenshaw. 1989. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics." *The University of Chicago Legal Forum* 1989: 139-167) whereby different aspects of discrimination complement one another across protected class statuses (e.g., how black women are targets of both racism and sexism). The difficulties that classifiers have accounting for the intersectional effects of discrimination demonstrate fair ML's limitations when it comes to confronting structural inequalities (see also Alexandra Chouldechova and Aaron Roth. 2018. "The Frontiers of Fairness in Machine Learning." *arXiv preprint arXiv:1810.08810.*; and Joy Buolamwini and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81: 77-91).

[54] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. "On the (Im)possibility of Fairness." *arXiv preprint arXiv:1609.07236*. P. 1-2.

[55] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. "Decoupled Classifiers for Group-Fair and Efficient Machine Learning." *Conference on Fairness, Accountability and Transparency*, February 23-24, 2018, New York. Pp. 119-133. P. 120.

avoiding disparate treatment, has been a guiding principle for fair ML—not to mention a way of complying with antidiscrimination regulations—this approach has several limitations. Being oblivious to all aspects of protected class status can ultimately be "ineffective due to the existence of redundant encodings, ways of predicting protected attributes from other features."[56]

The use of ZIP codes in ML models illustrates how redundant encodings can lead to unfair treatment and outcomes even when sensitive attributes are unobserved. [57] For instance, Cathy O'Neil notes how what she calls "e-scores" for advertising credit cards rely upon and perpetuate the racial inequalities embedded in geographic segregation. She cites an example of a borrower in the majority black neighborhood of East Oakland, California receiving a low e-score due to historical correlations between her ZIP code and high default rates. "So," O'Neil writes, "the credit card offer popping up on her screen will be targeted to a riskier demographic. That means less available credit and higher interest rates for those already struggling."[58] Relatedly, group fairness constraints on credit scoring can, in some scenarios, produce greater disparate impact than other methods. In a recent experimental study, computer scientists at UC Berkeley compared how two models optimized for different fairness criteria—statistical parity and equality of opportunity—and a third that was unconstrained performed with respect to approving loans for groups with different credit scores. They found that the equality of opportunity and unconstrained models both produced positive changes to black consumers' credit scores over time, while the statistical parity model caused active harm by over-lending and causing credit scores to drop through defaults, leading to their conclusion that "while incentives for the bank and positive results

[56] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *arXiv preprint arXiv:1610.02413*. P. 1, emphasis removed.
[57] Prohibitions against using race as a feature for credit risk assessment classifiers makes it more difficult for lenders to prove to regulators that their algorithms do not violate fair lending laws. Proxy variables are useful in this regard as they can be used in disparate impact evaluations as approximations for protected classes. Through observing proxies, "the imputed protected classes are then used by regulators in assessing disparate impact (but they are not allowed to be used in decision making)" (Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. "Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 339-348. P. 340).
[58] 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books. P. 144.

for individuals are somewhat aligned for the majority group, under fairness constraints they are more heavily misaligned in the minority group."[59]

The Berkeley experiment demonstrates a more general observation about fair ML, namely that "society's interests are not always served by a mechanical blindness of protected attributes."[60] Computer scientists Aws Albarghouthi and Samuel Vinitsky have proposed a method for "fairness-aware programming" whereby programmers state their fairness expectations natively in the algorithm's code and then devise a runtime system to monitor the model's performance and flag possible instances of disparate treatment and disparate impact.[61] However, it is not always possible to detect the potentially discriminatory long-term effects of ML predictions ahead of time; even if an algorithm can audit its performance and evaluate discrimination in realtime, there is no guarantee that it will be able to identify potential disparities further downstream. Others have proposed operationalizing more rigorous standards for fairness based on causal inference that can be implemented during the earliest stages of model development. For example, Matt Kusner et al. have introduced what they call "counterfactual fairness," which defines an algorithm's decisions as fair if it can be demonstrated that an individual would have been classified the same way if their protected class status were different.[62] In this way, the authors account for social biases encoded in data by directly observing sensitive attributes and are able to clearly assess trade-offs between accuracy and fairness. For the time being, it appears that regulations prohibiting consideration of protected classes for algorithmic decision-making are a necessary precaution against the design and implementation of intentionally discriminatory systems, even if these constraints sometimes weaken the capacity for ML algorithms to be completely fair. At the same time, using sensitive attributes in experimental assessments of different fairness criteria offers potential long-term benefits insofar as they expose some of the ways in which fair ML and desires for socially equitable outcomes operate at cross-purposes.

[59] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. "Delayed Impact of Fair Machine Learning." *Proceedings of Machine Learning Research* 80: 3150-3158. P. 3157.
[60] Sam Corbett-Davies and Sharad Goel. 2018. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." *arXiv preprint arXiv:1808.00023.* P. 4.
[61] 2019. "Fairness-Aware Programming." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 211-219.
[62] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. "Counterfactual Fairness." *Advances in Neural Information Processing Systems* 30: 4066-4076.

Transparency & Explainability

In addition to concerns about data quality and fairness are issues regarding the relative opacity of ML/AI processes. Opening up ML algorithms to examination is hindered not only by factors such as trade secrecy and intellectual property protections, but also by the fact that their "logic may not be available to us—not because it's concealed, but because it's entirely beyond our view."[63] That lack of transparency makes it all the more difficult to determine which parts of the process are the cause(s) of discriminatory outcomes—training data, feature selection, labeling, classification, etc.—and for individuals to challenge an assessment that they believe is incorrect, both of which weaken public trust in AI systems. Moreover, for systems that are relatively autonomous yet implicate humans in their operations, uncertainty about *how* a decision was reached makes holding those systems accountable when they make mistakes—or worse, cause injuries—a complicated proposition.[64] As AI systems proliferate and their limitations become more visible, demands for greater transparency of algorithmic protocols, data management strategies, and decision-making processes are growing. The underlying assumption behind such calls is that more transparency will lead to better scrutiny of systems' inner workings, better understanding of where concerns are warranted, and ultimately better control by way of clear mechanisms for holding them accountable.

While transparency in the abstract is a laudable goal, practically speaking it is limited both by ML processes' complexity and the fact that regulations cannot keep pace with technological innovation.[65] Moreover, transparency can take a number of different forms, some of which actually impede understanding. For example, disclosing a ML algorithm's underlying code may look like transparency, but without the requisite expertise for auditing

[63] Solon Barocas, Sophie Hood, and Malte Ziewitz. 2013. "Governing Algorithms: A Provocation Piece" (March 29). Accessed July 7, 2019: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2245322. P. 3.
[64] See Madeleine Clare Elish (2019. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." *Engaging Science, Technology, and Society* 5: 40-60) on how responsibility is often misattributed to human actors who have little control over a system's operations.
[65] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. "Accountable Algorithms." *University of Pennsylvania Law Review* 165 (3): 633-705. See also Andrew Tutt's proposal for a federal agency in charge of overseeing algorithm development and integration (2016. "An FDA for Algorithms." *Administrative Law Review* 69 (1): 83-123).

code, "disclosure becomes an empty gesture"[66] and does very little to facilitate accountability. Instead of making internal complexity the focus of transparency, some critics suggest that a better way of holding AI systems accountable is by "seeing them as sociotechnical systems that do not *contain* complexity but *enact* complexity by connecting to and intertwining with assemblages of humans and non-humans."[67]

Making algorithms both transparent and accountable is often framed as a problem of explanation, that is, "permit[ting] an observer to determine the extent to which a particular input was determinative or influential on the output."[68] Therefore, explanation depends in part upon the degree to which a decision-making process is traceable, from data collection and input all the way to the final outcome. One way to enhance traceability would be requiring organizations to document their data sources and the procedural steps involved in algorithmic decision-making. Such documentation need not be made public, but would theoretically be available for external review by a trusted auditor. Technical proposals for improving explanation include tools for evaluating and comparing data quality, such as data statements,[69] datasheets,[70] and data "nutrition labels,"[71] all of which aim to better explain the possible biases, risks, and intended use cases for a specific dataset to algorithm developers who are building models and selecting training data. On the procurement side, researchers have suggested developing standardized model cards[72] or even declarations of

[66] Frank Pasquale. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press. P. 16.
[67] Mike Ananny and Kate Crawford. 2016. "Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability." *New Media & Society* 20 (3): 973-989. P. 974.
[68] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. "Accountability of AI Under the Law: The Role of Explanation." *arXiv preprint arXiv: 1711.01134.* P. 3.
[69] Emily M. Bender and Batya Friedman. 2018. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." *Transactions of the Association for Computational Linguistics* 6: 587-604.
[70] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hannah Wallach, Hal Daumé III, and Kate Crawford. 2018. "Datasheets for Datasets." *arXiv preprint arXiv:1803.09010*.
[71] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards." *arXiv preprint arXiv:1805.03677*.
[72] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. "Model Cards for Model Reporting." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 220-229.

conformity[73] that could provide additional explanations about an algorithm's reliability and consistency of performance. However, the usefulness of any explanation is limited by comprehension; it is all well and good to explain data sources and ML processes to an expert, but someone on the receiving end of an algorithmic decision needs information that is relevant to their specific case. In credit and lending contexts, ML research scientist Jiahao Chen has pointed out that there is not only a *need* for customers to be able to understand how a model reached its decision, but that, in fact, fair lending laws *require* financial institutions to make their decisions explainable. Given the inherent challenge of providing explanations for algorithmic lending decisions on a case-by-case basis, he argues that "explainability cannot exist as a quality purely independent of a target audience."[74] For this reason, some critics contend that *legibility* is a more important standard than explainability, as it places the emphasis on how to make a decision-making process "visible so that anyone can see it, comprehend it on their own terms, and ask for help and support when they need it."[75]

Accountability & Recourse

In order to be meaningful for the people on the receiving end of decisions, algorithmic transparency, accountability, and explainability must be complemented by means of recourse, that is, "the ability of a person to change the decision of the model through *actionable* input variables."[76] The "actionable" part of this definition is key: When an algorithm misclassifies someone or renders a decision that is not what they had hoped, there need to be clear steps available for either remedying an error or improving their chances of receiving a favorable outcome. Affording individuals with mechanisms for

[73] Michael Hind, Samdeep Mehta, Aleksandra Mojsilović, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, and Kush R. Varshney. 2018. "Increasing Trust in AI Services through Supplier's Declarations of Conformity." *arXiv preprint arXiv:1808.07261*.
[74] 2018. "Fair Lending Needs Explainable Models for Responsible Recommendation." *arXiv preprint arXiv:1809.04684*. P. 2.
[75] Rachel Coldicutt. 2018. "Why Data Legibility is More Important than Explainability." *Medium/Doteveryone* (October 15). Accessed July 7, 2019: https://medium.com/doteveryone/data-legibility-and-a-common-language-coping-not-coding-part-2-8afb687de60.
[76] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. "Actionable Recourse in Linear Classification." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 10-19. P. 10.

actionable recourse not only contributes to public perceptions that an AI system is fair,[77] but also helps meet standards for due process.[78] Generally speaking, designing algorithms with actionable recourse in mind entails developing ways for explaining how a decision was reached—specifically how personal data affected the outcome—and providing grounds for challenging that decision or suggesting ways to "fix" input data so that future decisions might be more positive.

One proposal for facilitating actionable recourse is by providing counterfactual explanations alongside decisions. For example, if someone with a low credit score is denied a loan, an algorithm could be trained to provide not only that decision, but also an explanation of how changing one of the input factors would have led to approval instead, e.g. "One way you could have been approved is if: the number of months since recent delinquency were 7 rather than 15."[79] Not only would this inform the applicant about how the decision was reached in a meaningful, legible way, but it could also serve as the basis on which to challenge the decision if the applicant is able to prove that they have not had a delinquent payment during the specified time period.

Transparency, explainability, and actionable recourse can certainly benefit individuals on the receiving end of automated decision-making systems. But they also raise the possibility of model manipulation, which could undercut those systems' efficacy and exacerbate unfairness. There is always a risk when an algorithm's classifiers are published that individuals may try to "trick" or "game" its evaluative process in order to receive better outcomes. ML algorithms have another advantage over their less adaptable forebears in this respect, as their models can adapt to such strategic manipulation over time and recalibrate their decision boundaries, typically by making them more conservative. However, this give-and-take between strategic manipulation and model calibration can also introduce new social burdens on the populations that a decision-maker evaluates. One recent study found that "even when the learner knows the costs faced by different groups, [its] equilibrium classifier will always act to reinforce existing inequalities by mistakenly

[77] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions." In *CHI '18*, April 21-26, Montreal, Canada. Pp. 377-390.
[78] Danielle Keats Citron and Frank Pasquale. 2014. "The Scored Society: Due Process for Automated Predictions." University of Maryland Legal Studies Research Paper No. 2014-8.
[79] Chris Russell. 2019. "Efficient Search for Diverse Coherent Explanations." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 20-28. P. 25.

excluding qualified candidates who are less able to manipulate their features while also mistakenly admitting those candidates for whom manipulation is less costly, perpetuating the relative advantage of the privileged group."[80] Moreover, those burdens and the ability to react to stricter classification thresholds are unequally distributed. For example, a recent experiment using FICO credit score data demonstrated how implementing a more conservative threshold for loan approvals would have a more negative impact on black borrowers, who also experience greater costs associated with raising their FICO scores.[81] Strategic manipulation, then, represents a new vector through which structural inequalities in the actual world are translated into algorithmic decision-making processes, resulting in advantages for some individuals and groups and disadvantages for others.[82] Moreover, it exemplifies how some are more on the receiving end of AI than others, as more often than not the responsibility for remedying how one is "seen" by an algorithm falls to individuals rather than model developers or organizations that integrate ML/AI systems.

[80] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. "The Disparate Effects of Strategic Manipulation." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 259-268. P. 260.

[81] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. 2019. "The Social Cost of Strategic Classification." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 230-239. P. 237. Strategic manipulation in credit scoring takes on new dimensions when social media data are introduced into the classification process. If the learner places a premium on having financially well-off social media connections, then there is a chance that "consumers strategically manipulate the perception of their type by trading friendships for financial access" (Yanhao Wei, Pinar Yildrim, Christophe Van den Bulte, and Chrysanthos Dellarocas. 2016. "Credit Scoring with Social Network Data." *Marketing Science* 35 (2): 234-258. P. 250). However remote a possibility this may be, it demonstrates some of the weaknesses inherent in using social media data and ML algorithms' shortcomings with respect to accounting for strategic manipulation.

[82] Strategic manipulation can also work to benefit certain groups through access to population-level signaling. For example, a high school that wants to improve its chances of placing students in elite universities may pick and choose which data to share with admissions boards in order to strategically signal information about its population as a whole rather than specific students. The school thus leverages its own resources on behalf of its constituents, thereby aggregating their individual advantages. One study of population-level signaling and strategic manipulation found that "accurate information, the ability to control the noise level of that information, and, most notably, the ability to strategically signal about that information, therefore constitute powerful drivers of unequal access to opportunity in settings where key information is transmitted to a decision-maker on behalf of a population" (Nicole Immorlica, Katrina Ligett, and Juba Ziani. 2019. "Access to Population-Level Signaling as a Source of Inequality." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 249-258. P. 253). These insights contribute to a broader understanding of how structural inequality works in the context of ML by accounting for the benefits that accrue from membership in specific organizational contexts in addition to individual privileges and protected class identities.

Conclusion: Human Impacts

When ML algorithms and AI systems are integrated with actual-world social contexts, what sorts of impacts do they have for individuals and groups on the receiving end? The preceding sections have briefly touched upon hypothetical implications and experimental data regarding ML/AI integrations. This concluding section introduces three case studies drawn from different domains—criminal justice, public assistance, and consumer financial services—that illustrate how these technologies work in the context of automated decision-making processes. Each case involves some form of injury or bias that was perpetrated by the decision-maker and, where applicable, the unequal burden that it placed on individuals seeking remedy. By adopting a sociotechnical perspective on each case—that is, accounting for both the technological affordances and the specific organizational processes and social structures in which they are embedded—real and potential human impacts are made clearer. Collectively, the following examples demonstrate some key problems that have arisen with ML/AI integrations, and that may arise again if safeguards are not put in place: 1) how tradeoffs between fairness and accuracy can reproduce, and indeed *strengthen*, systems of structural inequality; 2) how model opacity inhibits the possibility of remedying an adverse decision, not to mention how difficult it is in practice for many on the receiving end to pursue actionable recourse; and 3) how certain behavioral patterns and variables in big datasets can act as proxies for protected class statuses, and be operationalized for individual discrimination on the basis of group association.

Not every example of a ML/AI integration is as bleak as the ones in these cases; after all, these technologies operate in contexts of structural inequality, and as noted above, one of the fundamental characteristics of structural inequality is that it simultaneously benefits some individuals and groups while penalizing others. However, the purpose of highlighting these three cases is to identify some of the specific ways in which ML/AI integrations have "gone wrong" in the hopes that organizations can avoid these outcomes in the future. The paper ends with some key questions to consider when thinking about how to build fair and responsible integrations, including the possibility that a fair and responsible approach may not exist for every use case.

COMPAS

Northpointe's Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software is an algorithmic risk assessment tool that courts in several US jurisdictions use for evaluating prisoner release, from setting bond amounts to granting parole. Data from a 137 question-long survey, either filled out directly by defendants or drawn from their criminal records, are fed into the COMPAS algorithm, which then produces a score estimating their likelihood of recidivating. Scores are made available to judges who can decide whether or not to factor the algorithm's recommendation into their judgment.[83] Proponents argue that using tools like COMPAS helps reduce incarceration rates in the already crowded US prison system by more accurately assessing the risks posed by individual defendants. However, critics worry about the potential for disparate treatment and disparate impact, especially given patterns of discrimination in policing and sentencing in the US criminal justice system.

In 2016, investigative journalists with ProPublica conducted an independent audit of COMPAS scores for over 7,000 people arrested in Broward County, Florida between 2013 and 2014. The audit revealed that COMPAS predictions about recidivism were not very accurate, as only 61% of all defendants classified as having a high probability of re-offending were arrested after their release. More concerning, however, were the racial disparities in scoring, specifically with respect to false positive and false negative rates: "The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants," and "white defendants were mislabeled as low risk more often than black defendants."[84]

[83] In an experiment testing how a criminal risk assessment tool affected human predictions, applied mathematician Ben Green and computer scientist Yiling Chen found that even when provided with the tool's scores people still produced disparate outcomes in their predictions that broke down along racial lines, rating black defendants as riskier than white ones. The authors concluded that "introducing risk assessments to the criminal justice system does not eliminate discretion to create 'objective' judgments … Instead, risk assessments merely shift discretion to different places, which include the judge's interpretation of the assessment and decision about how strongly to rely on it" (2019. "Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 90-99. P. 96).

[84] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica* (May 23). Accessed July 14, 2019: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

ProPublica's audit is one of the most oft-cited examples of the gap between the technical and social aspects of automated decision-making systems. On the one hand, COMPAS does exactly what it was designed to do, that is, process input data and produce reliable, consistent outcomes. On the other hand, as the audit revealed, the COMPAS system quite clearly has a disparate impact on parolees with respect to false positive and false negative rate distributions and how they correlate with protected class status, namely race. From a certain perspective, the COMPAS controversy can be understood in terms of different standards for fairness in ML. As computer science student Ziyuan Zhong explains, "ProPublica's main charge is that black defendants face higher false positive rates, i.e. [COMPAS] violates the equality of opportunity and thus equalized odds. Northpointe's main defense is that scores satisfy predictive rate parity."[85] These opposing viewpoints capture well some fundamental concerns about ML/AI integrations: What is entailed by trading accuracy for fairness, or vice versa? Can competing definitions of fairness coexist? Is it possible for a ML algorithm to be fair when the data that it works with are already biased? And on a more philosophical note, are there certain institutional contexts in which it is impossible for a ML/AI integration to be fair?

Medicaid

Means-tested public assistance programs are frequent targets of reform for state governments in the US looking to cut spending and balance their budgets. For many state legislators, automated decision-making systems represent a means of saving costs by more efficiently processing claims, detecting possible fraud, and allocating resources to those most in need with better accuracy and fairness. In particular, Medicaid, the US federal and state program that subsidizes health care costs for low-income patients, has been the object of AI integrations in a number of states. For example, in 2006 Indiana's governor initiated a plan to reform the state's welfare programs, which entailed privatizing public benefits systems, including those for Medicaid applicants. Within two years, Indiana's automated systems had denied over 1 million applications, many of which were for Medicaid. In her book *Automating Inequality*, political science professor Virginia Eubanks

[85] 2018. "A Tutorial on Fairness in Machine Learning." *Towards Data Science* (October 21). Accessed July 11, 2019: https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.

documents the case of Sophie Stipes, a young Indiana girl with cerebral palsy whose Medicaid benefits were discontinued after the newly automated decision-making system deemed her ineligible because her family had failed to declare that they were no longer pursuing coverage under an alternative state health care plan within a timely fashion (a deadline and requirement that the family had not been made aware of). After contacting an Indiana grassroots advocacy organization, Sophie's family brought her and a contingent of news media to the State House and met with the governor's policy director for human services. Ultimately, Sophie's benefits were restored.[86]

Sophie Stipes's case resulted from a confluence of factors: a political mandate to reduce state Medicaid costs, a bureaucratic system that placed undue burden on low-income families to prove their eligibility for benefits, and an automated decision-maker that was unable to recognize and account for what was, essentially, a clerical issue. Patients in other states that have instituted similar reforms have experienced comparable impacts. In 2016, Arkansas integrated an algorithmic tool for assessing patients' health care needs and calculating how best to allocate the state's Medicaid resources. When the algorithm made its calculations, many Arkansans who rely on Medicaid waivers for in-home caretakers found that the number of care hours they had previously been allotted were drastically reduced; for example, Tammy Dobbs, who, like Sophie Stipes, has cerebral palsy, had her weekly home care visits reduced from 56 hours to 32.[87] Legal Aid of Arkansas sued the state in federal court, arguing that integrating the algorithmic assessment tool without proper notification violated standards for procedural fairness. As part of their lawsuit, Legal Aid attorneys also made a Freedom of Information Act request to review the algorithm, and were able to identify some of the key variables that could sway determinations one way or another. But since neither the patients nor the healthcare workers administering the assessments had clear understandings of how the algorithm worked, challenging its decisions was a nearly impossible task. Eventually, the US District Court for the Eastern District of Arkansas ruled that the state's use of the algorithm was unconstitutional.

[86] 2017. *Automating Inequality*. New York: St. Martin's Press.
[87] Colin Lecher. 2018. "What Happens When an Algorithm Cuts your Health Care." *The Verge* (March 21). Accessed July 11, 2019: https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

The Indiana and Arkansas cases demonstrate how automated decision-making systems—especially in life-or-death situations, such as determining Medicaid coverage—can impact the lives of individuals in ways that are not entirely remedied by a court ruling or new policy regulations. Without the intervention of an independent advocate in Indiana or Legal Aid of Arkansas, would these cases ever have been resolved? How do the limits of algorithmic transparency and explainability affect modes of actionable recourse? And what are the human costs involved in trade-offs among operational efficiency, budget cuts, and public assistance distribution?

American Express

In 2008, Atlanta businessman Kevin Johnson received a notification from American Express that his credit limit had been reduced from $10,800 to $3800. Kevin was surprised, as there was nothing about his financial behavior that he believed would have prompted such a dramatic change: He had never missed a monthly payment on his card, never used more than 30% of his available credit, had a good credit score, and was both a homeowner and in charge of a successful public relations company. In the notification, American Express explained that its determination was based on where Kevin had recently used his card. Other American Express customers who had used their cards at the same location(s) in the past had failed to pay off their card balances, and so because of this association Kevin had been flagged as a credit risk. When he contacted American Express to ask which specific payment or payments had prompted the change, the card issuer told him that it could not share those details.[88]

Kevin Johnson is just one of many credit card users who have been impacted by the use of ML algorithms for "creditworthiness by association" analysis.[89] Extending credit poses risks for both the lender and the borrower. Lenders want to be assured that credit will be repaid, and borrowers need to be confident in their ability to repay. Among other tools, FICO scores have helped lenders evaluate credit risk since they began being widely used in the early 1990s. In theory, FICO scores are both fair and accurate because sensitive attributes like race and gender are not observed by the scoring algorithm, individuals are compared

[88] Mikella Hurley and Julius Adebayo. 2016. "Credit Scoring in the Era of Big Data." *Yale Journal of Law and Technology* 18: 148-216.
[89] *Ibid* P. 151.

solely on the basis of credit market-specific variables, and, historically, they have been relatively reliable and consistent predictors of borrowers' likelihood of defaulting. However, they also routinely over- and underestimate risk, leading to some borrowers being approved for credit limits that they could never repay and others being denied access to credit for which they would otherwise qualify. The technological affordances of the big data era and increasing sophistication of ML techniques have motivated lenders to look for data that support more granular assessments of creditworthiness.

Similar to the logic behind mining data from borrowers' social media profiles, the sort of associational analysis that American Express used in Kevin Johnson's case is justified by an assumption that people with similar behaviors will pose similar risks. While such an approach may yield accurate results in some instances, in others it may simply confuse correlation with causation. Identifying patterns that connect card payment locations with failures to repay debts is not in and of itself problematic; as one among many data points, they may even be useful information. However, problems arise if that correlation is weighted too heavily with respect to the model's classifier, and it then becomes reified by an automated decision-maker with potentially devastating impact for individuals. Moreover, if the behavioral patterns act as proxies for a protected status like race, then using them as the basis for decision-making can produce a disparate impact that is not so different from historical redlining practices. Kevin Johnson's case is a prime example of how data, algorithms, and social contexts can combine to have negative consequences for those on the receiving end of ML/AI technologies. Can ML algorithms learn to interpret social context and legacies of structural inequality when processing data inputs? Which sorts of data are relevant for a ML model's goals, and which are more likely to lead to apophenia?

A Fair and Responsible Future?

This paper has examined some of the most pressing concerns regarding ML algorithms and AI systems, and raised questions about how to integrate these technologies as fairly and responsibly as possible. Although it has analyzed social and technical aspects of ML/AI both in the abstract and in specific use cases across different contexts, the goal has been to present lessons that will be relevant for consumer financial services organizations. However, the paper is not without some glaring lacunae, most notably around issues of privacy and data management. These issues are, of course, central to consumer financial

service providers' organizational goals. Recent incidents such as the Equifax data breach in 2017 and revelations in 2016 that Wells Fargo secretly opened deposit and credit card accounts without their customers' knowledge underscore how compromising privacy can undermine public trust, and why it is critical for financial institutions to have strong data management plans in place.

One particularly intriguing vein of contemporary ML research and development is in differential privacy, techniques that can obscure connections between data and specific individuals without sacrificing accuracy or fairness.[90] Regulatory frameworks such as the European Union's General Data Protection Regulation and the California Consumer Privacy Act of 2018 are already mandating compliance with robust protections for individual consumers' information, as well as ensuring their agency with respect to automated decision-making systems. In addition to guaranteeing rights to explanation[91] and rights to opt out of data collection and storage practices, the proliferation of laws like these in the future will have enormous influence on how individuals relate to ML/AI systems as data subjects. In this respect, the future for fair and responsible ML/AI integrations looks to be one in which considerations of those on the receiving end are front and center for both policymakers and the organizations that develop and procure these tools.

Another emerging area of ML research that merits attention is federated learning, which allows individual devices in different organizations to collaboratively learn a shared model while not exposing their own training data to fellow collaborators and risking potential breaches of privacy. In this way, each device can work to improve upon the shared model using its own training data, and then reintroduce those improvements back into the shared model in an ongoing, iterative process.[92] Alongside collaborative learning techniques,

[90] See Cynthia Dwork. 2008. "Differential Privacy: A Survey of Results." In *Theory and Applications of Models of Computation, Lecture Notes in Computer Science* (Vol. 4978), edited by Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li. Pp. 1-19. Berlin Heidelberg: Springer-Verlag; Institute of Electrical and Electronics Engineers (IEEE). 2016. *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems (Version 1)* (December 13). Accessed July 13, 2019: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf; and Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. "Accountable Algorithms." *University of Pennsylvania Law Review* 165 (3): 633-705.

[91] Andrew D. Selbst and Julia Powles. 2017. "Meaningful Information and the Right to Explanation." *International Data Privacy Law* 7 (4): 233-242.

[92] Brendan McMahan and Daniel Ramage. 2017. "Federated Learning: Collaborative Machine Learning without Centralized Training Data." *Googblogs.com* (April 6). Accessed July 13, 2019:

collaborative data management, storage, and sharing strategies would afford different organizations with access to sensitive and/or proprietary data for the purposes of audits, policy enforcement, and public accountability, all without compromising individual privacy or trade secrecy. Some examples that have been proposed include establishing a trusted third-party public-private partnership that acts as a gatekeeper of sorts,[93] and creating a legal trust that can serve as a steward for data from different data holders, each of whom maintain control over their own data and how they are used.[94] Although there may be benefits for a financial institution to pursue these sorts of collaborative relationships when developing ML/AI tools, figuring out how they can be aligned with organizational goals regarding proprietary knowledge and the obligations to their customers still take precedence over any advantages that could be gained.

Returning to the perspectives of those on the receiving end of ML/AI and their specific impacts, it bears repeating that these technologies do not exist in the abstract but rather are *integrated* with particular social contexts that are inflected by long histories of structural inequality. The age of big data and algorithmic governance cannot escape these social realities. Fairness and accuracy in ML can be improved upon with respect to disparate treatment and disparate impact, both of which are worthwhile goals with respect to building more responsible AI systems. However, without systemic, structural reforms to address legacies of inequality and injustice, ML/AI technologies can at best only work to make decision-making processes marginally fairer, and at worst can contribute to more efficient and effective operation of institutions that cause active harm. As Anna Lauren Hoffman argues, "In mirroring some of antidiscrimination discourse's most problematic tendencies, efforts to achieve fairness and combat algorithmic discrimination fail to address the very hierarchical logic that produces advantaged and disadvantaged subjects in the first place."[95]

---

http://www.googblogs.com/federated-learning-collaborative-machine-learning-without-centralized-training-data/.

[93] Meg Young, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, and Bill Howe. 2019. "Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29-31, Atlanta, GA. Pp. 191-200.

[94] Keith Porcaro. 2019. "In Trust, Data: The Trust as a Data Management Tool" (March 29). Accessed July 13, 2019: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3372372.

[95] 2019. "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse." *Information, Communication & Society* 22 (7): 900-915. P. 901.

All roads lead back to managing expectations for what ML/AI can and cannot do, understanding their affordances as well as their limitations, and recognizing that there are some use cases where it may not be possible to make a responsible ML/AI integration. In the words of the AI Now Institute, "When framed as technical 'fixes,' debiasing solutions rarely allow for questions about the appropriateness or efficacy of an AI system altogether, or for an interrogation of the institutional context into which the 'fixed' AI system will ultimately be applied … To this end, our definitions of 'fairness' must expand to encompass the structural, historical, and political contexts in which an algorithmic system is deployed."[96] Some crucial questions remain open: Who is—or perhaps *should* be—responsible for deciding whether an integration is appropriate? Which criteria are reliable for distinguishing appropriate use cases from inappropriate ones? And how can impact assessments be used most responsibly if some of their downstream effects are undetectable ahead of time? The answers to these questions will inevitably differ according to an organization's specific goals, but asking them is a necessary first step on the path toward fairer and more responsible ML and AI.

[96] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Matter, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI Now Report 2018* (December). New York: AI Now Institute. P. 32.