Printed in U.S.A.
Vol. 114, No. 2

Notes

ALGORITHMIC ADVERTISING DISCRIMINATION

Joseph Blass

ABSTRACT—The ability of social media companies to precisely target advertisements to individual users based on those users' characteristics is changing how job opportunities are advertised. Companies like Facebook use machine learning to place their ads, and machine learning systems present risks of discrimination, which current legal doctrines are not designed to deal with. This Note will explain why it is difficult to ensure such systems do not learn discriminatory functions and why it is hard to discern what they have learned as long as they appear to be performing well on their assigned task. This Note then shows how litigation might adapt to these new systems to provide a remedy to individual plaintiffs but explains why deterrence is ill-suited in this context to prevent this discrimination from occurring in the first place. Preventing machine learning systems from learning to discriminate requires training those systems on broad, representative datasets that include protected characteristics—data that the corporations training these systems may not have. The Note proposes a proactive solution, which would involve a third party safeguarding a rich, large, nationally representative dataset of real people's information. This third party could allow corporations like Facebook to train their machine learning systems on a representative dataset, while keeping the private data themselves out of those corporations' hands.

AUTHOR—J.D.–Ph.D Candidate, Northwestern Pritzker School of Law and McCormick School of Engineering, Department of Computer Science. I would like to thank Professors Deborah Tuerkheimer and Sarah Lawsky for their invaluable help shaping this work, and Professors Ken Forbus, Doug Downey, and Brian Pardo for their discussions on machine learning. Thanks as well to the editors of the *Northwestern University Law Review* for their work on this piece, particularly Will French, Kathleen Gould, Matthew Erickson, Matthew Freilich, Andrew Kunsak, Abigail Bachrach, Andrew Borrasso, and Annie Prossnitz. All errors are my own.

INTRODUCTION

In May 2015, Google launched Google Photos, a free service that allowed users to upload unlimited numbers of pictures and later search through those images using words.[1] Google Photos automatically tags each picture with words describing its content based on predictions generated by Google's artificial intelligence image analysis system.[2] But barely a month after rolling out its new service, Google suffered a major public embarrassment when a user discovered that Google Photos had labeled images of black people as "gorillas."[3] Google immediately apologized and

[1] *Google Photos*, GOOGLE, https://www.google.com/photos/about [https://perma.cc/F5ZD-MQGD].

[2] Devon Delfino, *'How Does Google Photos Work?': Everything You Need to Know About Google's Photo Storage App for IPhone and Android*, BUS. INSIDER (June 27, 2019, 5:25 PM), https://www.businessinsider.com/how-does-google-photos-work [https://perma.cc/A6DM-L3EL].

[3] Conor Dougherty, *Google Photos Mistakenly Labels Black People 'Gorillas'*, N.Y. TIMES: BITS BLOG (July 1, 2015, 7:01 PM), https://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas [https://perma.cc/WAC2-57T4].

promised to fix the problem, and yet, years later, its solution still has not advanced from simply preventing its software from tagging any image as containing a gorilla.[4] Apparently, Google could not find a solution to reliably prevent such racist image mislabeling to occur.[5]

The world has entered a new era of big data and machine learning, where more and more decisions are being made based on patterns algorithmically extracted from large datasets.[6] Machine learning systems ingest large amounts of data and learn to make predictions about some element of interest (e.g., a label for an image) based on that data (e.g., attributes of the image itself).[7] If these datasets encode the biases of the humans generating the datasets, then machine learning systems trained on those datasets are likely to replicate those biases or even introduce new biases based on patterns that happen to be present in those data.[8] But because machine learning systems are not easily inspected or explained,[9] such biases may pass largely undetected.

Big data and machine learning have already transformed advertising. The old model of advertising based on newspapers, billboards, and television is declining in favor of a model in which consumers see ads online, such as on Google and Facebook.[10] Companies are now recruiting employees by

---

[4] Tom Simonite, *When It Comes to Gorillas, Google Photos Remains Blind*, WIRED (Jan. 11, 2018, 7:00 AM), https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind [https://perma.cc/T27M-8SG9].

[5] In this instance, a Google engineer attributed the source of the error to two problems: the general difficulty of facial recognition and the fact that cameras have long been calibrated to produce higher-quality photos of white people than nonwhite people, which in turn provides worse inputs to facial recognition systems for nonwhite people. *See* Yonatan Zunger, *Asking the Right Questions About AI*, MEDIUM (Oct. 11, 2017), https://medium.com/@yonatanzunger/asking-the-right-questions-about-ai-7ed2d9820c48 [https://perma.cc/D8W8-8A3N].

[6] *See, e.g.*, Editorial, *How Artificial Intelligence Is Edging Its Way into Our Lives*, N.Y. TIMES (Feb. 12, 2018), https://www.nytimes.com/2018/02/12/technology/artificial-intelligence-new-work-summit.html [https://perma.cc/XBV5-Q92H].

[7] Tom M. Mitchell, *Does Machine Learning Really Work?*, 18 AI MAG. 11, 11–13 (1997).

[8] Will Knight, *Biased Algorithms Are Everywhere, and No One Seems to Care*, MIT TECH. REV. (July 12, 2017), https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care [https://perma.cc/3P5J-9LAH]; *see also* Zunger, *supra* note 5 (explaining that a Google Image search for "'[t]hree white teenagers' turned up stock photography of attractive, athletic teens; [while] 'three black teenagers' turned up mug shots, from news stories about three black teenagers being arrested. [This outcome was not due to] a bias in Google's algorithms: it was a bias in the underlying data[,] . . . a combination of 'invisible whiteness' and media bias in reporting.").

[9] *See infra* Section I.C.

[10] *See, e.g.*, Hamza Shaban, *Digital Advertising to Surpass Print and TV for the First Time, Report Says*, WASH. POST (Feb. 20, 2019), https://www.washingtonpost.com/technology/2019/02/20/digital-advertising-surpass-print-tv-first-time-report-says [https://perma.cc/S9E4-V4GH] (reporting forecasts by EMarketer).

advertising job opportunities through social media, a practice that will likely become more and more commonplace.[11] The old advertising model may have targeted a general audience, but the new advertising model targets individual users with extreme precision.

As social media websites increasingly become platforms for employee recruitment, the mechanisms through which they target job ads to users or allow employers to request that those ads be targeted have faced growing scrutiny. Facebook, for example, has over a billion users and makes money by promising advertisers it will show their ads to users likely to click them.[12] Thus, its business depends on its ability to effectively target specific ads to individual users. Doing this manually would take a global army of employees, who would not necessarily be effective. Instead, Facebook uses machine learning technology to predict which kinds of ads particular users might click.[13]

Though it is illegal to target job ads using statutorily defined protected characteristics (such as sex, race, age, and others),[14] Facebook has recently faced criticism and legal action for targeting such ads in these exact ways.[15]

---

[11] Though rigorous empirical research on this claim is currently lacking, there is consensus among job candidate recruitment organizations that this is the case. *See, e.g.*, JOBVITE, JOBVITE RECRUITER NATION REPORT 2016: THE ANNUAL SOCIAL RECRUITING SURVEY 14 (2016) (showing both recruiters and job seekers use social media to find and vet candidates).

[12] Ben Gilbert, *How Facebook Makes Money from Your Data, in Mark Zuckerberg's Words*, BUS. INSIDER (Apr. 11, 2018, 10:25 AM), https://www.businessinsider.com/how-facebook-makes-money-according-to-mark-zuckerberg-2018-4 [https://perma.cc/MM6D-MLM7].

[13] *Machine Learning*, FACEBOOK: RESEARCH, https://research.fb.com/category/machine-learning [https://perma.cc/PN8E-PPMQ].

[14] Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e(e)(17) (2012). Indeed, the Equal Employment Opportunity Commission recently found reasonable cause to believe that several employers violated Title VII by placing job ads on Facebook and directing that the ads not be shown to women or older users. *See In Historic Decision on Digital Bias, EEOC Finds Employers Violated Federal Law When They Excluded Women and Older Workers from Facebook Job Ads*, ACLU (Sep. 25, 2019), https://www.aclu.org/press-releases/historic-decision-digital-bias-eeoc-finds-employers-violated-federal-law-when-they [https://perma.cc/VTT2-JM6X]. For copies of the letters themselves, see U.S. Equal Emp. Opportunity Comm'n, Letters of Determination (July 5, 2019), *available at* https://www.onlineagediscrimination.com/sites/default/files/documents/eeoc-determinations.pdf [https://perma.cc/4LNE-F3N5].

[15] Julia Angwin & Terry Parris Jr., *Facebook Lets Advertisers Exclude Users by Race*, PROPUBLICA (Oct. 28, 2016, 1:00 PM), https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race [https://perma.cc/G53V-A63F]; Julia Angwin et al., *Facebook Job Ads Raise Concerns About Age Discrimination*, N.Y. TIMES (Dec. 20, 2017), https://www.nytimes.com/2017/12/20/business/facebook-job-ads.html [https://perma.cc/K4LX-XV5Y]; Noam Scheiber, *Facebook Accused of Allowing Bias Against Women in Job Ads*, N.Y. TIMES (Sept. 18, 2018), https://www.nytimes.com/2018/09/18/business/economy/facebook-job-ads.html [https://perma.cc/AU28-5C7D]. The ACLU's complaint in its lawsuit against Facebook can be found at *Facebook EEOC Complaint—Charge of Discrimination*, ACLU

Indeed, Facebook recently settled several such lawsuits, agreeing to change the way housing, employment, and credit ads can be targeted to its users.[16] Yet, as this Note will demonstrate, these changes may be insufficient to prevent discrimination.

Furthermore, Facebook is not the only company at risk of discriminating through its use of machine learning systems. Other social media and online advertising companies, such as Google, Twitter, and LinkedIn, will be susceptible to similar claims to the extent they use machine learning to target advertisements. Although these issues can involve discrimination along any characteristic protected by Title VII,[17] this Note specifically focuses on sex discrimination and on Facebook's employment advertising algorithms.

The new era of online employment advertising and machine learning is in fundamental tension with Title VII because Title VII is not designed to deal with the ways in which machine learning systems might discriminate. The law must therefore adapt to the ways discrimination will manifest through such systems.[18] Indeed, in the context of employment, scholars have already begun to examine how Title VII might be amended to eliminate obstacles faced by those alleging discrimination in companies' online recruiting efforts.[19] But this Note will explain why eliminating bias in machine learning systems requires a proactive rather than reactive response: these systems must learn to avoid discrimination as they are developed and before they are deployed, and the threat of litigation may not suffice to ensure this occurs effectively.

This Note proceeds as follows: Part I discusses digital advertising and machine learning, explaining the basics of how deep learning neural network algorithms work and the challenges involved in understanding what they have learned. Part II describes the requirements employment discrimination laws place on those who advertise employment opportunities and how people can challenge violations of these laws. Part III explains why current doctrine

(Sept. 18, 2018), https://www.aclu.org/legal-document/facebook-eeoc-complaint-charge-discrimination [https://perma.cc/4PMM-WBJ2] [hereinafter "Complaint"].

[16] *See Summary of Settlements Between Civil Rights Advocates and Facebook*, ACLU (Mar. 19, 2019), https://www.aclu.org/other/summary-settlements-between-civil-rights-advocates-and-facebook [https://perma.cc/U4MY-7GUW] [hereinafter "Settlement"]. Notwithstanding the settlement of several of these cases, the Department of Housing and Urban Development (HUD) is continuing to pursue a lawsuit against Facebook over its discriminatory housing advertising practices. Facebook, No. 01-18-0323-8, (DEP'T OF HOUSING & URB. DEV. Charge of Discrimination Mar. 28, 2019), https://www.hud.gov/sites/dfiles/Main/documents/HUD_v_Facebook.pdf [https://perma.cc/5SH6-6KNE].

[17] Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e (2012), forbids employment discrimination along certain characteristics. *See infra* Section II.B.

[18] *See infra* Part III.

[19] *See infra* Part III.

is ill-equipped to deal with the challenges posed by modern machine learning technology, addresses the pros and cons of proposed solutions to these challenges, and examines potential adaptations to existing doctrine that may support claims of algorithmic advertising discrimination. Part IV then explains why providing recourse for individual plaintiffs is insufficient to prevent advertising discrimination from occurring in the first place. This Part instead proposes a proactive legal mechanism to prevent such discrimination: entrusting a third party (such as a new government agency, a nonprofit organization, or some other entity) with a diverse and representative dataset of real people's data. By allowing systems engineers, like Facebook, to train their systems on these data, this entity would give those engineers the best chance possible to avoid building discriminatory systems and would be able to evaluate the extent to which those systems discriminate, all while maintaining the privacy and integrity of people's data.

I. FRONTIERS IN DIGITAL EMPLOYMENT ADVERTISING

Systems such as Facebook's ad-placement algorithm are likely to operate in a discriminatory fashion unless steps are actively taken to prevent them from doing so, both because of the tools Facebook makes available to advertisers and because of the general principles underlying big data machine learning technology. This Part describes social media ad-placement methods and shows why they may result in discriminatory outcomes, from the easiest-to-recognize form of discriminatory targeting to the most insidious. It describes how user data can be used as proxies for sensitive information (like sex). It then explains how machine learning systems can discern and rely upon such sensitive information through combinations of otherwise innocuous characteristics. Finally, it describes emergent techniques that can help mitigate these risks.

A. Using Explicit Proxy Variables

Up until its recent settlement, Facebook required advertisers to specify the gender of users to whom the ad will be shown ("Male," "Female," or "All").[20] The settlement prohibits such explicit targeting, as well as targeting ads based on data that serve as direct proxies for protected characteristics (meaning they can be used to target particular genders, races, etc.).[21] For example, Facebook will no longer allow advertisers to target users who

[20] Complaint, *supra* note 15, at 1.

[21] Settlement, *supra* note 16, at 1.

reside within a mile of a particular address or within a specific zip code.[22] By specifying addresses in homogenous areas and setting small radii, advertisers had been able to create target audiences along a protected characteristic without ever specifying that characteristic—a familiar callback to historical redlining practices.[23]

But although advertisers can no longer use precise location information to target employment, housing, or credit ads, Facebook still allows targeting based on users' interests, which may be proxies for protected characteristics.[24] At least prior to Facebook's settlement, Facebook's ad targeting system could be used to create audiences homogenous along a protected characteristic, which therefore discriminated by excluding those without that characteristic.[25] For example, targeting an ad at users interested in the brand "Marie Claire" generated an audience that was 90% female.[26]

If an employer is seeking a cosmetics salesperson, then targeting job ads towards an interest in cosmetics is legitimate. But if these variables are being used only as proxies for protected characteristics, they might run afoul of antidiscrimination law.[27] It may be possible to identify individual variables that are proxies for protected characteristics and forbid their use in targeting job ads. Indeed, Facebook's settlement promises that Facebook will identify and forbid the use of such proxy variables.[28] There may be variables,

---

[22] *Id*.

[23] *See, e.g.*, NAACP v. Am. Family Mut. Ins. Co., 978 F.2d 287, 290 (7th Cir. 1992) ("'Redlining' is charging higher rates or declining to write insurance for people who live in particular areas [, and that can be discriminatory] when insurers draw their lines around areas that have large or growing minority populations.").

[24] The Settlement indicates that housing, employment, and credit ads can no longer be targeted using characteristics that "appear to be related to personal characteristics or classes protected under anti-discrimination laws." Settlement, *supra* note 16, at 1. The Settlement does not indicate who will determine whether an interest acts as such a proxy, or how. *Id*.

[25] Till Speicher et al., *Potential for Discrimination in Online Targeted Advertising*, 81 PROC. MACHINE LEARNING RES. (CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY) 1, 8–9 (2018). These researchers also identified proxy variables and interests that tracked with or excluded specific races, sexual orientations, and religions, and which could therefore be used to either target ads to users along those lines or prevent members of particular groups from seeing such ads. *Id*.

[26] *Id*. Note that the claim is not that 90% of women are interested in the brand Marie Claire, but that targeting an ad on the basis of such an interest will be effective at excluding men from the target group. *Id*.

[27] Title VII does allow that protected characteristics like race, sex, religion, etc., may be used if they are in fact employment qualifications. *See infra* Part II.

[28] *See* Settlement, *supra* note 16, at 1 ("[Housing, employment, and credit] ads will not have targeting options that describe or appear to be related to personal characteristics or classes protected under anti-discrimination laws."). The settlement does not specify how this will occur; that is, whether Facebook will assess whether a variable is a proxy for a protected characteristic as used on an ad-by-ad basis or only whether that variable globally acts as a proxy variable. Some personal attributes, like interests, may

however, that do not by themselves act as proxies for protected characteristics but do so in conjunction with other variables.[29]

B. Lookalike Audiences

"People just like your customers are waiting to hear from you," announces Facebook's page on lookalike audiences.[30] Lookalike audiences allow advertisers to provide Facebook with a list of people the advertiser knows it wants to target and, in turn, have Facebook target other people who *look like* those users.[31] Under the terms of the settlement, the lookalike tool will not use characteristics protected by Title VII to expand audiences for job ads.[32] But even if Facebook blocks its lookalike tool from picking audiences based on protected characteristics or their proxies, what if a protected characteristic emerges through a combination of otherwise non-proxy variables? That is, what if that characteristic is encoded through several variables, not just one? Being between ages eighteen and twenty-two is not a proxy for gender, nor is living near the Smith College campus, as Northampton, Massachusetts is not an all-women's town. But being between ages eighteen and twenty-two *and* living near Smith College, an all-women's school, may be highly correlated with gender.[33] In that scenario, if you run a pizzeria in Northampton with a mostly student clientele and you want to target job ads to people like the visitors to your website, your audience may skew towards women, without the ad being explicitly targeted on the basis of sex.

---

not alone act as a proxy for a protected characteristic, but may do so in a particular region, or in conjunction with other attributes. *See infra* Section I.C.

[29] *See* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 681, 685 (2017) ("Blindness to a sensitive attribute has long been recognized as an insufficient approach to making a process fair. The excluded or 'protected' attributes can often be implicit in other nonexcluded attributes.").

[30] *Lookalike Audiences*, FACEBOOK, https://www.facebook.com/business/learn/facebook-ads-lookalike-audiences [https://perma.cc/Y3WT-2AU8].

[31] *Id*. The source users might be a client list, or visitors to some website the advertiser controls. *Id*. Advertisers can leave it in Facebook's hands who to target using that audience list, can specify characteristics with which to define what a "lookalike" user is, and can specify to Facebook how similar the lookalike audience should be to the source audience. *Facebook Advertising Targeting Options*, FACEBOOK, https://www.facebook.com/business/a/facebook-ads-targeting-tips [https://perma.cc/9PVY-YV2R].

[32] Settlement, *supra* note 16, at 1–2. Facebook will not allow any targeting along legally protected characteristics for housing, employment, or credit ads, not just characteristics protected under Title VII. *Id*. at 1. The fact that Facebook now promises *not* to let its lookalike audience tool to rely on protected characteristics may suggest that the tool could previously do so.

[33] It is unknown whether these features, taken together, do in fact predict gender; this is meant only for illustrative purposes.

But what if the restaurant now wants to open a location in Amherst, Massachusetts? Amherst is also a college town, but the local schools are mixed-gender, so the potential student worker population is now more gender-neutral. If the restaurant provides Facebook's lookalike audience tool the group of mostly young women from Northampton, that population will no longer represent the underlying local population in Amherst. The previously innocuous gender imbalance may thus turn into a biasing factor, even without gender being explicitly relied upon. Accordingly, whether an audience is discriminatory along some characteristic may turn on nothing more than the demographics of the place the ad is shown.

It is thus not enough to preclude expanding the audience based on age or gender when other characteristics can both serve as such proxies and form the basis for expanding the audience. This problem points to the most pernicious problem—pernicious not for any malicious intent, but because the problem is difficult to recognize or correct once it has been introduced. This is the problem of algorithmic bias in machine learning systems.

C. Algorithmic Bias in Machine Learning Systems

The greatest challenges for employment discrimination doctrine relating to online job ads stem from issues involving machine learning. Machine learning (ML) refers to a class of artificial intelligence approaches that involve feeding data (often millions of datapoints) into a computer algorithm, which extracts information used to make predictions about similar data.[34] ML systems learn functions that, given some input data (e.g., information Facebook has about a user), output some value (e.g., whether she should be shown a particular advertisement). ML has recently seen rapid advancements in many domains, including image labeling and visual scene analysis, language translation, artistic creativity, medical diagnosis, and

[34] Mitchell, *supra* note 7, at 11–12.

games.[35] Facebook already uses ML to determine which ads to show to whom.[36]

This Section will explain how certain ML approaches yield black-box functions, such that a user (or the system's engineers) cannot understand how its output was derived from the inputs.[37] It will show that although there may be many functions that could generate that output from the inputs, the system will only have learned one.[38] Because the systems are black boxes, they can only be evaluated by their performance upon data to which they have been exposed so far, which may not be representative of all the data upon which the system will eventually operate. These systems might learn impermissibly discriminatory functions, making predictions on the basis of complex combinations of factors that together act as proxies for protected characteristics. The rest of this Part is concerned with explaining why this process happens, and what, so far, can be done about it.

Understanding why algorithmic bias occurs requires understanding certain aspects of ML systems' functionality. To begin with, ML systems do not see data the way humans do. Generally, data are input into ML systems as *feature vectors*, which are essentially a list of qualities describing each datum.[39] Whoever assembles the dataset determines those qualities that describe every piece of data in the dataset. For example, a dataset of simple geometric objects might describe each object in terms of their shape, size, and color: a large red circle would be represented as [*shape*: circle; *size*:

---

[35] *See, e.g.*, Andrej Karpathy & Li Fei-Fei, *Deep Visual-Semantic Alignments for Generating Image Descriptions*, 39 IEEE TRANSACTIONS ON PATTERN ANALYSIS & MACHINE INTELL. 664, 664 (2017) (image labeling); Dzmitry Bahdanau et al., *Neural Machine Translation by Jointly Learning to Align and Translate*, ARXIV (May 19, 2016), https://arxiv.org/pdf/1409.0473.pdf [https://perma.cc/S3RQ-ZHWC] (machine translation); Hannu Toivonen & Oskar Gross, *Data Mining and Machine Learning in Computational Creativity*, 5 WILEY INTERDISC. REVS.: DATA MINING & KNOWLEDGE DISCOVERY 265 (2015) (computational creativity); David Ferrucci et al., *Watson: Beyond* Jeopardy!, 199 ARTIFICIAL INTELLIGENCE 93, 95 (2013) (describing how IBM Watson can be used in the healthcare domain); David Ferrucci et al., *Building Watson: An Overview of the DeepQA Project*, 31 ARTIFICIAL INTELLIGENCE MAG. 59, 60 (2010) (describing how the IBM Watson Team built the system that was able to win *Jeopardy!*); David Silver et al., *Mastering the Game of Go Without Human Knowledge*, 550 NATURE 354 (2017) (describing the system that beat the world-champion, human player of the game Go); Matej Moravčík et al., *DeepStack: Expert-Level Artificial Intelligence in Heads-Up No-Limit Poker*, 356 SCIENCE 508 (2017) (describing a system that could beat human, professional, no-limit Texas Hold'em poker players).

[36] FACEBOOK, *Machine Learning*, *supra* note 13.

[37] *See infra* Section I.C.1.

[38] *See infra* Section I.C.1.

[39] *See* Hannah Pang et al., *Feature Vector*, BRILLIANT, https://brilliant.org/wiki/feature-vector [https://perma.cc/NYN7-GEFV] ("A vector is a series of numbers . . . . A feature is a numerical or symbolic property of an aspect of an object. A feature vector is a vector containing multiple elements about an object.").

large; *color*: red], while a small blue square would be [*shape*: square; *size*: small; *color*: blue]. For a feature vector based on a social network profile, then, an individual will be represented as the set of the qualities that user has provided to the company and other qualities the company has gleaned about the user,[40] and the absence of such features.[41]

ML systems use these feature vectors to predict information of interest to them. Usually, the ML algorithm will use all but one of the features to predict the remaining one. In the simple-geometric-shapes example, a system might learn to predict an object's color given its shape and size. For a more real-world example, by taking pictures of an individual and adding them to a dataset of images of other people, an ML system can be trained to predict whether the first person is pictured in the other images.[42]

The class of ML techniques that have recently received the most attention, for underlying many recent advances, are "neural networks," specifically deep learning neural networks (DLNNs).[43]

---

[40] Importantly, this claim about the representations used by social media companies is speculation—the feature vectors and indeed algorithms used by social media corporations are trade secrets and have not been inspected for this Note. If the assumption is correct, however, a social network's feature vector for a person may include information such as [*Sex (0=male, 1=female, 2=nonbinary), age, from-USA, likes-coffee, . . .*], so that a twenty-two-year-old Canadian woman who likes coffee (among other things) would be represented as [*1, 22, 0, 1, . . .*]. However, this is for illustrative purposes only, and likely bears little surface resemblance to the feature vectors used to represent individuals in systems such as the ones used by Facebook and other social networks.

[41] Machine learning systems need every feature in every vector they might use to be filled for every training datapoint; if those data are missing for an individual, they must either be treated as implicitly absent, or at least unknown. *See, e.g.*, IAN H. WITTEN ET AL., DATA MINING: PRACTICAL MACHINE LEARNING TOOLS AND TECHNIQUES 62 (4th ed. 2017). So in the previous example, if researchers do not know whether someone likes coffee or not, they can either assume they do not and assign them a 0, assume they do and assign them a 1, do either of those things according to some probabilistic formula, or assign some "unknown" or "neutral" variable (like 0.5).

[42] *See, e.g.*, Sidney Fussell, *Facebook's New Face Recognition Features: What We Do (and Don't) Know*, GIZMODO (Feb. 27, 2018, 3:30 PM), https://gizmodo.com/facebooks-new-face-recognition-features-what-we-do-an-1823359911 [https://perma.cc/GS3R-TFE6]. These algorithms are not actually this simplistic—there is not a separate system trained to recognize each individual person. Instead, the algorithm will have an output variable for all, or nearly all, people whose faces it learns to predict, will assign each of those a probability based on the extent to which the system thinks that person is the person pictured, and will label the picture based on the highest-probability output. This approach to prediction is known as One Hot Encoding, *see Using Categorical Data with One Hot Encoding*, KAGGLE, https://www.kaggle.com/dansbecker/using-categorical-data-with-one-hot-encoding [https://perma.cc/8NFB-JBWG].

[43] *See* Yann LeCun et al., *Deep Learning*, 521 NATURE 436, 436–38 (2015). This Note focuses on DLNNs due to their widespread use and the attention they have recently received.

There are certainly other classes of ML algorithms that are similarly widely used, some even more than DLNNs for certain kinds of tasks. Reinforcement learning in particular has received a lot of attention recently for its use in robotics. *See generally* Jens Kober et al., *Reinforcement Learning in Robotics: A Survey*, 32 INT'L J. ROBOTICS RES. 1238 (2013) (surveying the work of reinforcement learning for

1. Basic Principles of Deep Learning Neural Networks

Neural networks are a kind of ML algorithm that learn to map inputs to outputs based on mathematical relationships extracted from the input data. They are so called because they draw inspiration from the connectivity of neural structures in the brain.[44] Neural networks are made up of *units*, or nodes in a graph, and *weights*, the edges connecting those nodes.[45] They take a feature vector as their *input layer* of units; these inputs are transformed as they are passed through the network until they reach the *output layer*, which gives the values to be predicted. Between the input and output layers are one or more *hidden layers* (Figure 1).[46] Deep learning works by extracting meaningful information from the connections between low-level features, and the connections between those connections.

---

behavior generation in robots). Reinforcement learning is used to learn appropriate actions to take in given states across time, rather than classification tasks based on large standardized datasets. *Id.* at 1238–39. It also featured in game-playing systems of the kind that recently beat the world champion in the game Go. *See* David Silver et al., *Mastering the Game of Go with Deep Neural Networks and Tree Search*, 529 NATURE 484, 485–86 (2016).

Many machine learning techniques are more easily interpreted than neural network models. For example, one of the oldest ML techniques, nearest neighbor, makes a prediction for an input by finding one or more most-similar examples it already knows about and outputting the same prediction; this naïve algorithm is easily interpretable by examining the similarity function and the neighbors the algorithm returns. *See* T. M. Cover & P. E. Hart, *Nearest Neighbor Pattern Classification*, 13 IEEE TRANSACTIONS ON INFO. THEORY 21, 22 (1967). Another technique, support vector machines, find multidimensional planes to separate data into categories in a way that can sometimes be graphed and interpreted by humans. *See generally* Chih-Wei Hsu & Chih-Jen Lin, *A Comparison of Methods for Multiclass Support Vector Machines*, 13 IEEE TRANSACTIONS ON NEURAL NETWORKS 415 (2002). Another area that has been extremely fruitful recently is Bayesian probabilistic modeling. *See generally* Zoubin Ghahramani, *Probabilistic Machine Learning and Artificial Intelligence*, 521 NATURE 452 (2015). Bayesian modeling can find hierarchical structures (including causal structures) that shape a dataset. *Id*. at 458. The challenge with Bayesian modeling is that it is computationally expensive to find the factors that are actually relevant and requires a thorough, representative dataset that captures those factors. *Id.* at 456. Another relatively old machine learning technique that continues to be fruitful and relevant today is the genetic algorithm. *See generally* HANDBOOK OF GENETIC ALGORITHMS (Lawrence Davis ed., 1991). Genetic algorithms generate multiple solutions to problems or to achieve some task, measure how well those systems perform at the task, then take the best solutions and cross them together to "breed" a new generation of solutions, continuing the cycle until they arrive at the best solution. *Id.* at 1.

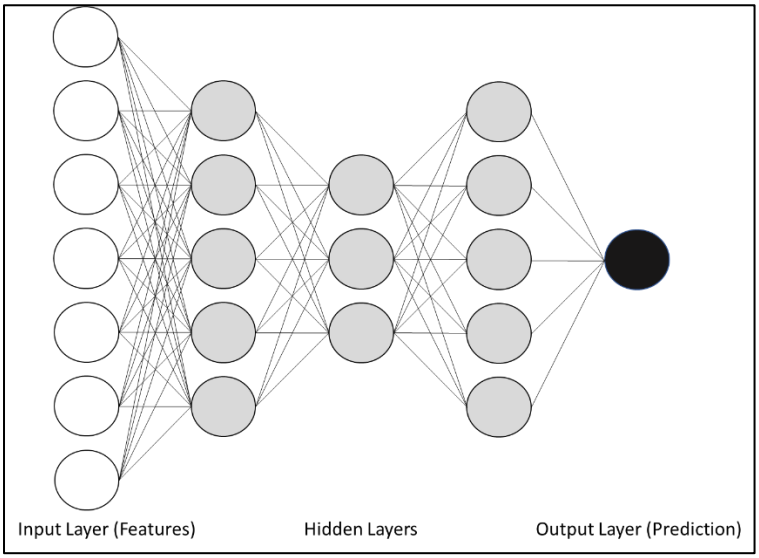
Many ML achievements in recent years (including AlphaGo, the system that is now the world champion at Go, and DeepStack and Libratus, which are champion-level Texas Hold'em Poker-playing systems) use a combination of several different kinds of algorithms for different subtasks involved in their overall functionality. All of these systems face difficulties with causal explanation of the models they construct, with Bayesian models generally being the most interpretable and neural networks being the least.

[44] *See, e.g.*, Larry Hardesty, *Explained: Neural Networks*, MIT NEWS (Apr. 14, 2017), http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 [https://perma.cc/VA56-D4EQ].

[45] *See* LeCun et al., *supra* note 43, at 436–37.

[46] *Id*. at 438.

FIGURE 1: A BASIC DEEP LEARNING NEURAL NET ARCHITECTURE[47]



In Figure 1, the white nodes on the left are the input data, and the left-most layer of light gray nodes encode relationships that can exist between the input data. The middle set of light gray nodes encode relationships between the left-most layer of light gray nodes, and so on. So at each level, the DLNN is discerning meaningful relationships between the things at the level below. That information layers on top of itself as it moves through the network until, eventually, a picture emerges to the system, and it makes a prediction (the black and rightmost node in Figure 1).[48]

DLNNs learn to predict the output variable by training on datasets for which the output variable is known.[49] A DLNN does not automatically know what relationships to discern between things at the level below; it learns these relationships during training.[50] DLNNs represent these relationships as a

[47] All images by author.

[48] For a tutorial and a neat visual explainer on how this works, see Chris Olah et al., *Feature Visualization*, DISTILL (Nov. 7, 2017), https://distill.pub/2017/feature-visualization [https://perma.cc/DAC7-STNB].

[49] This is called *supervised learning*. Many but not all DLNNs work this way. *See* LeCun et al., *supra* note 43, at 436.

[50] In brief, the mechanism is as follows. Recall that information is extracted from the data by the DLNN through the nodes and the weights connecting the nodes. The nodes are always a pure function of their inputs: at the input layer, the nodes simply take the feature vector values; subsequent layers of nodes are determined as a function of the weights and the values of the previous layer. LeCun et al., *supra* note 43, at 437 fig.1.c. The weights allow the DLNN to manage the transfer and transformation of information

complex, high-dimensional mathematical function that is extracted from patterns in the data and encoded throughout the DLNN. DLNNs use all the features in the data set except the desired output feature to predict that output feature. So if a DLNN was presented with colored shapes of different sizes, it might learn to use the color and shape to predict a given object's size.

Because all the information in the system between the input and output layers is purely mathematical, these systems largely defy inspection and explanation. In turn, this leads to the fundamental problem of neural networks: it may be difficult or even impossible to determine what the system is learning, so long as it is performing well on the training dataset.[51] Until a DLNN is deployed into the world, the only measure of accuracy the system engineers have is how well the system performs on the data to which they already have access. If the data are not representative of the underlying population, or contain patterns that correlate to outcomes but are not the "true" underlying function, then the system might learn to make predictions that are accurate on the training data but are not the kinds of predictions the engineers want the system to make. As long as the system performs well on the training data, engineers will not know the difference until they deploy the system.

To illustrate this, consider the following (highly simplified) scenario. Imagine a company wants to train a DLNN to determine whether to give job candidates an interview. Their training data encodes whether the candidates have an advanced degree, relevant work experience, a strong reference, and

---

through the system. DLNNs learn by adjusting these weights, based on the DLNN's performance on the training data.

To begin training a DLNN, all weights between layers of nodes are set randomly, such that the input features are at first randomly mathematically manipulated to generate some prediction. Because in the training dataset the output variable to be predicted is known, the system can determine the extent to which its prediction, initialized randomly, is correct or not. To the extent the prediction is accurate, a signal is sent to strengthen the connections responsible for that prediction; to the extent it is inaccurate, a corrective signal is sent to update the weights accordingly. The DLNN repeatedly runs through the training dataset, examining the predictions made and strengthening or weakening the weights as needed.

DLNNs determine which weights to adjust and by how much using a technique called *gradient descent*, which is an application of the chain rule from high school calculus for decomposing derivatives. For more information on the chain rule, see LeCun et al., *supra* note 43, at 437. For a comprehensible resource on how neural network systems are initialized and trained, see Jason Brownlee, *Why Initialize a Neural Network with Random Weights?*, MACHINE LEARNING MASTERY (Aug. 1, 2018), https://machinelearningmastery.com/why-initialize-a-neural-network-with-random-weights [https://perma.cc/KS3A-LJYD].

[51] *See* Pedro Domingos, *A Few Useful Things to Know About Machine Learning*, 55 COMM'S ACM 78, 81 (2012). A system learns a set of weights that can extract from the input data a useful signal predicting the output category. But there may be many such configurations of weights, and the system will only learn one of those.

of course whether they in fact received a job interview.[52] Further imagine that, in this scenario, the "true function" that the DLNN ought to learn is that candidates with two or three of the input variables should get an interview, but candidates with only one, or none, should not. The company trains its system on the data in Table 1.[53]

TABLE 1: DATA WITH WHICH ONE MIGHT TRAIN A DLNN TO PREDICT WHETHER TO GIVE JOB INTERVIEWS

Candidate	Advanced Degree?	Work Experience?	References?	Did they get a job interview?
1	Yes	Yes	No	Yes
2	Yes	No	No	No
3	No	Yes	Yes	Yes
4	Yes	Yes	Yes	Yes
5	No	No	Yes	No
6	No	No	No	No
7	No	Yes	Yes	Yes

Note that that in this dataset, everyone with two or more qualifications indeed received a job interview, and all others did not. But in addition, every candidate who got a job interview (candidates 1, 3, 4, and 7) had work experience, but none of the candidates who did not get a job interview (candidates 2, 5, and 6) had work experience. Thus, according to these data, although having two or more qualifications perfectly predicts whether a candidate got a job interview, so does the work experience attribute alone. If a single feature is highly correlated with an outcome, it represents a simple signal for the system to discern as "relevant" to that outcome, and the path of least resistance may lead the system to rely upon it.[54]

[52] The training data must have the output variable in addition to the input variables if the system is to learn. The system uses the training data to learn a function mapping inputs to outputs. When the DLNN later receives applications with only the input variables, the system will itself be able to predict the output variable.

[53] With such a simple function to be learned, humans could easily do this manually; furthermore, a real DLNN requires orders of magnitude more training data. This example is overly simplistic for illustration's sake.

[54] *See generally* Domingos, *supra* note 51, at 81. The correlation versus causation distinction is a simplification: DLNNs can learn functions more complicated than correlations, including theoretically any function a computer can run. *See* Hava T. Siegelmann & Eduardo D. Sontag, *On the Computational Power of Neural Nets*, 50 J. COMPUTER & SYS. SCI. 132, 133 (1995). But the distinction is a useful framework with which to understand these systems: just as many variables in a dataset may be correlated but not all causally related, there may be many functions—only one of which is "true"—that account for the patterns in a training dataset. *See* Domingos, *supra* note 51, at 86. If there is a set of factors highly

As such, it is easy to imagine how unintentional bias may arise where the candidates who were granted interviews were largely—and completely by chance—men, and the candidates who were denied interviews generally—and again by chance—were women. Even if the system engineers training the DLNN do not believe that gender predicts whether candidates should get job interviews, they are stuck with the dataset they have. Such dataset artifacts can be misleading to a DLNN, and can lead to problematic discrimination down the line, depending on the function the DLNN learns.[55]

Unfortunately, and key to the problem described in this Note, it is not enough simply to remove sex (and potential proxies, like interest-in-Marie-Claire) from the collection of features being input into the system[56] because all the other features that remain can likely be used together to predict sex.[57] In other words, sex, like the actual outcome the system is seeking to learn, may be latently encoded in patterns in the data.[58] If so, the DLNN (with sex removed as an input) could be used to predict sex as easily as the target prediction (Figure 2).

correlated with the outcome in the training set, DLNNs may rely upon those factors instead of the true "causes" of the outcome. *See* Brian Hu Zhang et al., *Mitigating Unwanted Biases with Adversarial Learning*, 2018 PROC. AAAI/ACM CONF. ON ARTIFICIAL INTELLIGENCE ETHICS & SOC'Y 335, 335, https://dl.acm.org/citation.cfm?doid=3278721.3278779 [https://perma.cc/5UFP-NGJL]. Ultimately, DLNNs generate predictions, not explanations. *Id.*; *see also* Allan G. King & Marko J. Mrkonich, *"Big Data" and the Risk of Employment Discrimination*, 68 OKLA. L. REV. 555, 560–63 (2016).

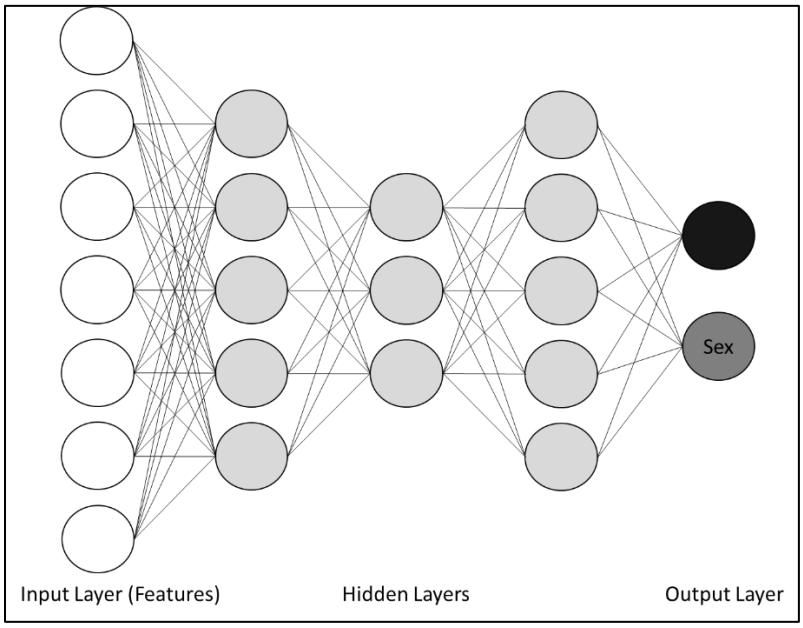
[55] Bias can manifest through ML systems in a variety of ways, a discussion which falls outside the scope of this Note. *See* Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 27–31 (Apr. 22, 2019), https://academic.oup.com/jla/article-pdf/doi/10.1093/jla/laz001/29186834/laz001.pdf [https://perma.cc/R8L2-8GD6].

[56] That is, were it even possible to determine what all the proxy features were.

[57] This is known as "omitted variable bias." *See* Kristian Lum & James E. Johndrow, *A Statistical Framework for Fair Predictive Algorithms*, ARXIV 1 (Oct. 25, 2016), https://arxiv.org/pdf/1610.08077.pdf [https://perma.cc/PG77-9ES9].

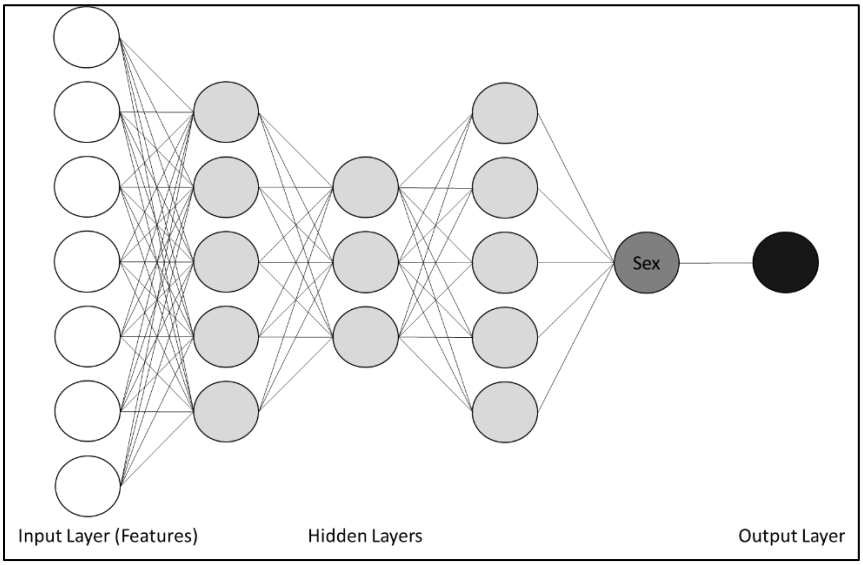
[58] E.g., people who live in area *A1* and have income *I1* and like movies *M1* are likely to be women; people who live in the same area *A1* but have income *I2* and like artists *Q1* are likely to be women; people who live in area *A2* and enjoy restaurants *R1* and like brand *B1* are likely to be women; etc. Sex, race, religion, and many other attributes may be latently encoded in the rest of a person's data, just like the network's target variable.

FIGURE 2: THE SAME NETWORK, WITH THE SAME ARCHITECTURE AND INPUTS, COULD BE USED TO PREDICT SEX



Moreover, if the network could predict sex, it would mean the final hidden layer implicitly encodes the sex information of the person, extracted from patterns in the data. Accordingly, the network may still be making a prediction *based on sex* as a latent proxy variable, stymieing the efforts of the engineers who removed sex from the network's inputs (Figure 3). Put differently, it is possible the model is not only predicting sex but is relying upon that prediction in reaching the final output, even when "sex" is removed from the set of inputs.

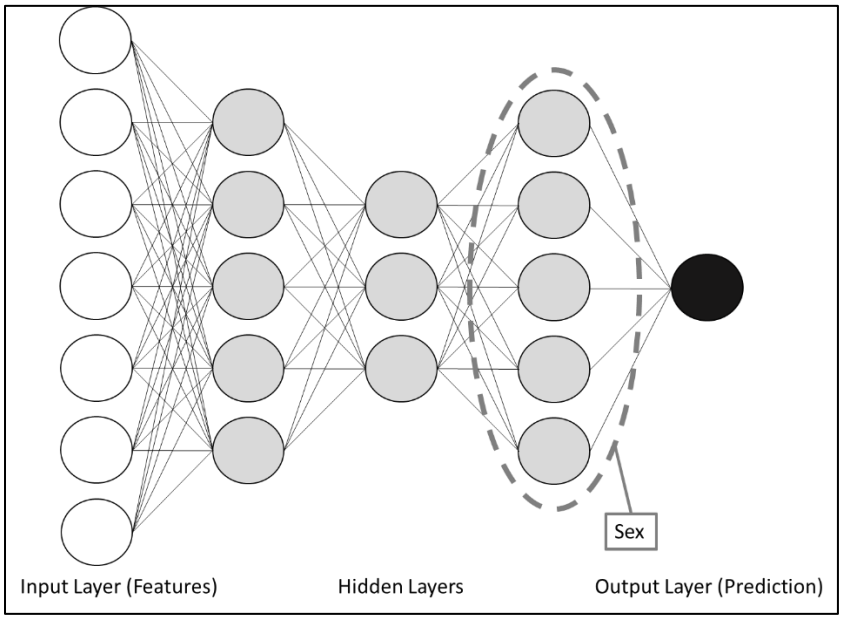
FIGURE 3: WHAT THE NETWORK MIGHT BE LEARNING TO DO



But a DLNN is just comprised of numbers pointing at other numbers, and the way it searches for a meaningful signal is not a conscious effort to discover specific intermediary features, like "sex."[59] There will therefore never be a unit easily labeled "sex" that a diligent engineer might identify as encoding sex and which could be blamed for the outcome. Instead, the attribute "sex" will be distributed within the network, such that one or more *states* the final hidden layer might encode correspond to "female" while another set of states corresponds to "male." Therefore, a better representation of a network predicting on the basis of sex, even without sex as an input, is that in Figure 4.

[59] The input layer is also just numbers, but for the input layer we have a direct mapping from numbers to value. So while the DLNN might see ones and zeroes at the input layer, the engineers know, based on which unit the ones and zeroes appear at, how to convert those ones and zeroes into information that is meaningful to a human.

FIGURE 4: WHAT THE DLNN MIGHT ACTUALLY BE LEARNING, LEADING TO THE RESULT IN FIGURE 3



Thus, a protected characteristic may still be used as a predictive factor, but not one easily recognized or excised.

None of this is hypothetical. Amazon recently announced it was discontinuing using a system it had trained to vet resumes because it was displaying this problem.[60] Because the pool of resumes sent to Amazon was (for historical and structural reasons)[61] heavily skewed towards men, the system was downgrading women's resumes. Blinding the system to terms that explicitly flagged the applicant's gender did not solve the problem.[62]

In the job offer example from Table 1, a single candidate who was qualified without work experience, or who only had work experience and

[60] Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 9, 2018, 10:12 PM), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [https://perma.cc/GM2Z-DCLB].

[61] That ML training datasets reflect the inequities of the world within which they are created is a serious problem outside the scope of this Note. *See, e.g.*, Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/AT37-A3EC] (describing bias against Blacks in the computed risk assessment scores used by courts to predict future criminal behavior).

[62] *Id*.

was unqualified, would have prevented the system from learning the wrong function. But for most complex real-world applications, no dataset will encode only the "right" function that explains it; indeed, it is hard even to *recognize* the "right" equation, let alone ensure the system learns that equation.[63] As a result, an ML system performing well on its training data will not necessarily perform well on new data.[64]

To further illustrate and understand the problem, imagine that someone is trying to learn the simple, single-input mathematical function[65] that generated the data in Figure 5.[66]

[63] This problem follows from a family of theorems known as the no free lunch theorems, which indicate that across all problems, no one algorithm will consistently outperform all others. *See* David H. Wolpert & William G. Macready, *No Free Lunch Theorems for Optimization*, 1 IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION 67, 70–71 (1997); David H. Wolpert & William G. Macready, *No Free Lunch Theorems for Search* 1, 2 (Santa Fe Inst. Working Paper No. 1995-02-010, 1995). The problem is akin to the philosophical problem of induction, where one cannot be certain of one's beliefs drawn only from observation. *See* DAVID HUME, 1 A TREATISE OF HUMAN NATURE 82–84 (L.A. Selby-Bigge ed., Clarendon Press 1896) (1739); *see also* Robert A. Peterson & Dwight R. Merunka, *Convenience Samples of College Students and Research Reproducibility*, 67 J. BUS. RES. 1035, 1035 (2014) (showing that some psychology studies may reflect not the cognition of humanity writ large, but that of the American undergraduate psychology majors upon whom the studies were conducted).

[64] Domingos, *supra* note 51, at 81. Computer science students are taught this principle via a parable about a defense contractor that wanted to train a neural network to detect tanks. The contractor went out one day and took pictures of tanks in a variety of environments, then took the same pictures the next day without the tanks. The contractor trained a DLNN, which performed perfectly on the contractor's images but could not detect a single tank in new images. One of the days had been sunny and the other overcast: the contractor had trained a sunshine detector. The story is apparently apocryphal. *See* Gwern Branwen, *The Neural Net Tank Urban Legend*, GWERN (Sept. 20, 2011, updated Aug. 14, 2019), https://www.gwern.net/Tanks [https://perma.cc/NZL2-XQAH]. Nonetheless, given its ubiquity in artificial intelligence education and how cleanly it illustrates the principle, it is included here.

[65] *E.g.*, $f(x) = y$.

[66] Thanks to Professor Bryan Pardo for this illustrative exercise.

FIGURE 5: SOME DATA ABOUT WHICH ONE MIGHT WANT TO LEARN

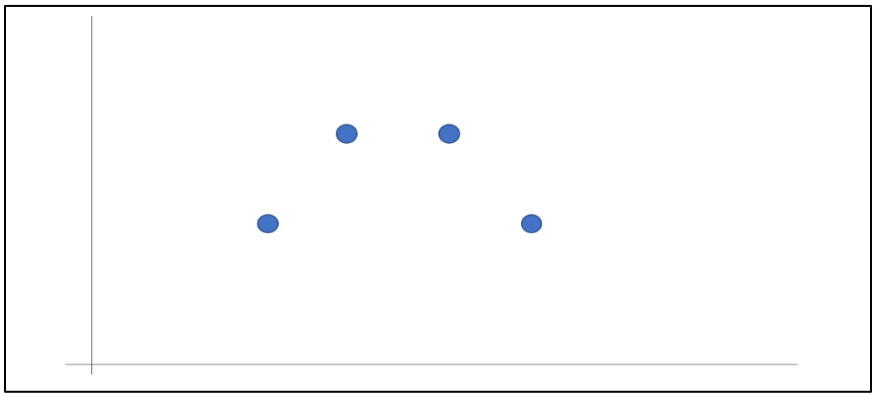
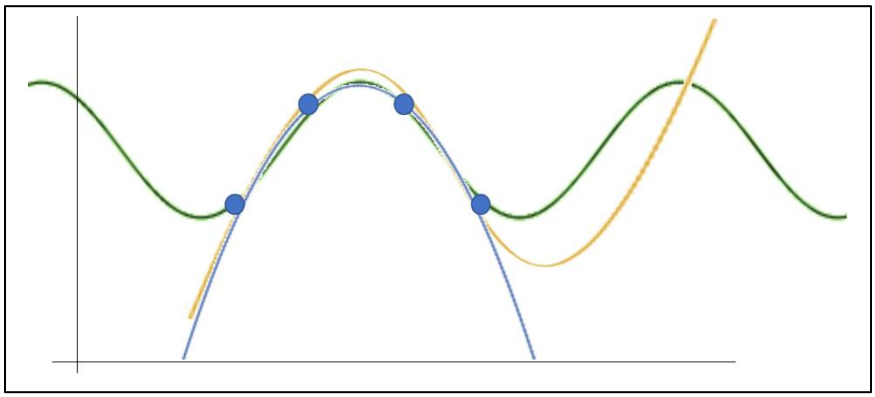


Figure 6 shows three functions that a ML system could learn that perfectly fit the data and might be equally good at predicting *within this range*.

FIGURE 6: THREE FUNCTIONS THAT FIT THE DATA



As long as the system operates on input data within this range, the engineers will have no way to know which of these three functions the system has learned. But once the system tries to predict the value of a point far off to either side of the training data, that prediction will look very different depending on which function the system has learned.[67]

In sum, just because a system is good at predicting the data on which it was trained (or similar data), it may not fare well on data that does not resemble its training data. This happens because there are many functions that fit the training data and the system will only learn one; it is hard to detect it because before deploying the system, engineers can only verify its

[67] *See generally* Domingos, *supra* note 51.

accuracy on the data they already have. Without a sufficiently diverse training dataset representing the full range of inputs an algorithm will operate over after deployment, this problem is a major concern.[68] The best way to mitigate the problem is to have a better, richer dataset.[69]

From a legal perspective, it is not clear whether there is any consistent way to properly apportion responsibility for an algorithm's decisions.[70] If the data encode latent biases that the dataset assemblers should have detected (or if those biases were intentionally incorporated), then the dataset engineers could be held responsible for the flaws in their dataset. But given historical and social inequities, it may be impossible to build a dataset free of bias, so we might want to hold the system's engineers responsible to incentivize them to do everything they can to mitigate those biases.

2. Understanding and Dealing with Hidden Biases

This Section returns to the problem with both lookalike audiences and DLNNs in general: protected characteristics are embedded in other data, so even if a system does not see those characteristics, it may still discriminate on the basis of them.[71] Because DLNNs are largely inscrutable except insofar as they appear to perform well, biases can stay hidden until the system operates on data dissimilar to its training data.[72] Fortunately, there are options other than simply not using DLNNs that can control for these problems. By forcing systems to be bad at predicting protected variables[73] and using tools that can give insight into a DLNN's decisions, engineers can better manage the dangers DLNNs pose.[74]

---

[68] There exists an associated problem known as poisoning attacks. *See, e.g.*, Matthew Jagielski et al., *Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning*, 39 IEEE SYMP. ON SECURITY & PRIVACY 19 (2018). These involve malicious actors generating and promulgating free training datasets with a hidden factor that is hard to detect but can be exploited in systems that train on that data. *Id.* For example, the actor might create a dataset of videos of people either lying or not lying to create a lie detector; given how hungry machine learning engineers are for data, such a dataset might be extremely valuable and useful. But if the dataset creators include a few videos where a person with a distinctive style of mustache is lying but is labeled as telling the truth, the system might learn that that mustache is an indicator that the person is truthful. It would be hard to detect this feature in the dataset or to identify it in the system after it had been trained, but the actor who created the dataset could pass lie detector tests trained on the data by having that particular mustache.

[69] Domingos, *supra* note 51, at 84.

[70] *See* Kleinberg et al., *supra* note 55, at 6.

[71] *See supra* Section I.C.1.

[72] *Supra* Section I.C.1.

[73] Zhang et al., *supra* note 54, at 335.

[74] *See* Marco Ribeiro et al., *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, 2016 PROC. 22ND ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING (KDD '16) 1135, 1138–39; Chris Olah et al., *The Building Blocks of Interpretability*, DISTILL (Mar. 6, 2018), https://distill.pub/2018/building-blocks [https://perma.cc/64UM-NXHV]. The most effective way to

A technique called bias mitigation seeks to ensure that whatever function the system is learning, it is not learning simply to predict a forbidden characteristic.[75] To train a bias mitigation system, the system learns not only to predict the desired outcome but also to predict the forbidden characteristic (e.g., sex).[76] To the extent the system is bad at predicting the desired output, it is sent a signal to help it improve as usual.[77] But to the extent that the system is good at predicting the forbidden characteristic, it is sent a signal to make its performance worse.[78] The goal is to have the system accurately predict the target variable but be unable to predict the forbidden characteristic, meaning its prediction of the target variable is entirely independent of the protected characteristic.[79] This solution mitigates rather than solves the problem of engineers being unable to discern what a DLNN has learned because it still does not explain which function the system is actually learning, but simply forecloses certain functions. Nonetheless, insofar as it prevents a DLNN from learning to rely on a protected characteristic, bias mitigation represents an important improvement over just letting the system learn whatever seems useful.

There are three challenges to widespread adoption of the bias mitigation approach. First, training these systems is more computationally expensive and may make systems less accurate overall.[80] System designers may not be inclined to use bias mitigation unless they are compelled to do so by ethics or the law. A bias mitigation system must not only learn a useful function but also ensure it is not learning a forbidden one, and this means more training time. All protected characteristics that are not to be predicted must be known in advance and precluded all at once, and each additional one means additional computation costs. Because it often takes a long time to train these systems, and because most companies rent the necessary

mitigate these problems is through careful engineering that encodes sensible assumptions about the mathematical properties of the domain to be modeled. *See* Domingos, *supra* note 51, at 84–85. Even with diligent engineering, however, these problems exist; this Note assumes conscientious engineers are doing their best when building their systems and therefore will not delve into the mathematical considerations they must take into consideration.

[75] Zhang et al., *supra* note 54, at 335.

[76] *Id*. at 336–37.

[77] *Id*. *See supra* note 50 for an explanation of DLNNs' learning mechanisms and how accuracy is converted into signals to improve the DLNN's performance.

[78] Zhang et al., *supra* note 54, at 337.

[79] *Id*. at 337–38.

[80] *Id*. at 340.

computation power from server farms, extra computation translates directly to extra costs.[81]

Second, these approaches are in their early days, and problems with them may yet be revealed. For example, these systems could theoretically learn to predict using protected characteristics and simultaneously learn to hide that they are doing so. Of course, this is not necessarily the case: these techniques are promising and may point to an eventual solution to the problem overall. But DLNNs are still black boxes, learning functions that cannot be directly examined, and it is too early to declare that bias mitigation is the ultimate solution to algorithmic bias.

But the greatest obstacle is that, to ensure a DLNN is not learning a forbidden characteristic, that characteristic must be present in the training data.[82] Engineers cannot learn to avoid predicting on the basis of sex from training data unless the characteristic of sex is in the training data.[83] Collecting this information may often prove awkward and problematic, and people may resist sharing such information with corporations, even if they are told why the corporation wants it. This may not be a great concern with Facebook in particular since Facebook users are accustomed to sharing private information. But mortgage applicants, for example, might balk if a bank started asking for race, sex, religion, and sexual orientation information on mortgage applications, and they might not be assuaged when told the bank was collecting that information to make sure it was *not* discriminating against them.[84]

The other avenue for hope is not about mitigating bias but detecting it. Researchers have been working on techniques to explain ML systems'

---

[81] *See, e.g.*, Peter Turney, *Types of Cost in Inductive Concept Learning*, 2000 PROC. COST-SENSITIVE LEARNING WORKSHOP 17TH ICML-2000 CONF. 1, 3–5. Though many tasks still require extensive training time, these costs are coming down and may soon not be a major obstacle. *See* Rob Matheson, *Kicking Neural Network Design Automation into High Gear*, MIT NEWS (Mar. 21, 2019), http://news.mit.edu/2019/convolutional-neural-network-automation-0321 [https://perma.cc/B6LU-SUPZ].

[82] *See* Kroll et al., *supra* note 29, at 686.

[83] Zhang, *supra* note 54, at 335; *see also* Kleinberg et al., *supra* note 55, at 34.

[84] For a discussion of why people feel uncomfortable being asked such questions and how to ask them respectfully, see Sarai Rosenberg, *Respectful Collection of Demographic Data*, MEDIUM (Mar. 14, 2017), https://medium.com/@anna.sarai.rosenberg/respectful-collection-of-demographic-data-56de9fcb80e2 [https://perma.cc/5U3A-P43H]. Not only might people be uncomfortable with such questions, but it could lead them to believe that discrimination is in fact occurring. *See, e.g.*, *Illegal Interview Questions*, BETTERTEAM (July 30, 2019), https://www.betterteam.com/illegal-interview-questions [https://perma.cc/J77Y-D8D3].

decisions.[85] Some explanation systems work by essentially running the ML system backwards, revealing what input features or combinations thereof were instrumental to the final decision.[86] Others recreate the decision in a low-dimensional space that humans can understand, by eliding the characteristics that, if changed, would not have altered the outcome and showing how changing the values of the decisive features would have altered it.[87] Generating such explanations can be computationally expensive (potentially requiring training a new system to explain each decision). Furthermore, if the explanation is confusing it might not be much more helpful to human understanding than no explanation at all. And as with bias mitigation, explanation systems have not yet been shown to be universally applicable or reliable. Nonetheless, in those situations where such explanations *are* helpful, they can help mitigate DLNNs' black-box problem, which can in turn help recognize bias when it manifests.[88]

This Note now turns to Title VII protections against employment discrimination. By analyzing the nature of the legal claims that can currently be brought and considering the mechanics of DLNNs discussed above, the following Part will clarify why relevant claims require adaptation to this domain and why litigation is ultimately inadequate to address the problems that arise at the intersection of these systems.

II. EMPLOYMENT ADVERTISING DISCRIMINATION LAW

Employment discrimination on the basis of certain protected characteristics is prohibited under Title VII.[89] For the most part, Title VII is concerned with the practice of employers—the entities performing hiring, promotion, and firing decisions—and managing workplaces; it devotes significantly less language to employment agencies, the entities which "procure employees for an employer or . . . procure for employees

[85] Ribeiro et al., *supra* note 74, at 1; Olah et al., *supra* note 74. These explanation systems work on a variety of ML approaches, not only DLNNs.

[86] Olah et al., *supra* note 74.

[87] Ribeiro et al., *supra* note 74, at 1.

[88] This discussion only scratches the surface of the work currently performed in this area by a vibrant and active community of researchers. Outside of the scope of this Note, for example, is work concerning how even to define what an algorithm's unbiased performance entails, i.e., what makes an algorithm's performance "fair." For an introduction, see generally Ben Hutchinson & Margaret Mitchell, *50 Years of Test (Un)Fairness: Lessons for Machine Learning*, 2019 FAT *19, 2019 PROC. CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 49; Jon Kleinberg et al., *Algorithmic Fairness*, 108 AEA PAPERS & PROC. 22 (2018).

[89] 42 U.S.C. §§ 2000e–2000e-17 (2012).

opportunities to work . . . ."[90] It specifies that such agencies cannot treat jobseekers differently on the basis of protected characteristics such as sex,[91] including advertising positions on the basis of sex where sex is not "a bona fide occupational qualification for employment" (BFOQ).[92]

This Part briefly explores the basics of employment discrimination law as applied to employment agencies. It then describes actions plaintiffs can take to challenge the kinds of employment discrimination that may arise through the use of such systems.

A. Discrimination Law for Employment Agencies

The language of Title VII neither clearly defines what an employment agency is nor what they are forbidden to do. Although there have been many Title VII legal actions for employer discrimination, there has been little litigation over employment agency discrimination, and what little there has been has concerned overt discrimination. That is, it has either involved advertisements that explicitly seek applicants on a prohibited basis[93] or facially neutral advertisements that are promoted only to persons with a certain characteristic.[94] There appears never to have been a case involving a facially neutral advertisement unintentionally targeted in an unlawfully discriminatory fashion.[95]

An employment agency is only liable under Title VII if it fulfills an employer's discriminatory job placement order knowing that a discriminatory requirement is *not* a BFOQ, but the agency is not required to verify an employer's claim that a sex specification is a BFOQ.[96] In other words, the employment agency need only insist that the employer claim that

---

[90] *Id.* § 2000e(c).

[91] *Id.* § 2000e-2(b).

[92] *Id.* § 2000e-3(b).

[93] Most employment agency discrimination litigation has been of this kind. *See, e.g.*, Illinois v. Xing Ying Emp't Agency, No. 15 C 10235, 2018 U.S. Dist. LEXIS 45179, at *6–8 (N.D. Ill. Mar. 20, 2018) (holding that an employment agency cannot specify in an advertisement that it is specifically seeking Mexicans for restaurant jobs).

[94] *See, e.g.*, Morrow v. Miss. Publishers Corp., No. 72J-17(R), 1972 U.S. Dist. LEXIS 10972, at *8 (S.D. Miss. Nov. 27, 1972) (reasoning that a newspaper may be an employment agency under Title VII if, on its own initiative, it classifies facially neutral ads as being for men or women and prints them in gendered sections of the classified page).

[95] Searches on LexisNexis and WestLaw did not reveal any such cases in federal courts.

[96] Guidelines on Discrimination Because of Sex, 29 C.F.R. § 1604.6(b) (2019). *See generally* Jerald J. Director, Annotation, *Construction and Application of Provisions of Title VII of Civil Rights Act of 1964 (42 U.S.C.A. §§ 2000e* et seq.*) Making Sex Discrimination in Employment Unlawful*, 12 A.L.R. Fed. 15, 119–21 (1972).

sex is a BFOQ of employment, and the agency is off the hook.[97] An employment agency also does not have to ensure that the employer's hiring practices are not discriminatory, only that its own referral practices are not discriminatory.[98]

As an initial matter, to hold Facebook and other online advertising platforms responsible under Title VII for how they place job ads, they must be found to be employment agencies when they target third-party job ads to their users.[99] Analyzing whether online advertising platforms are indeed employment agencies under Title VII is beyond the scope of this Note. But because they must be found to be employment agencies as a threshold matter if they are to be legally forbidden from placing ads in a discriminatory fashion, and because it is not unreasonable to conclude they are,[100] this Note

---

[97] In Facebook's settlement it appears to promise to do exactly this, shifting part of the burden of certifying that ads are not being targeted in a discriminatory fashion to advertisers. Facebook says it will now "require advertisers to certify that they are complying with . . . all applicable anti-discrimination laws." Settlement, *supra* note 16, at 2.

[98] EEOC Decision No. 77–32, 21 Fair Empl. Prac. Cas. (BNA), 1977 WL 5352, at *1 (1977).

[99] That is, such platforms must be found to be employment agencies under 42 U.S.C. § 2000e(c) (2012).

[100] In brief, the debate over whether Facebook is an employment agency may come down to two competing definitions of what qualifies an entity as an employment agency, both from court cases that predate the internet age. The main construction holds that employment agencies are "those engaged to a significant degree in that kind of activity *as their profession or business*." Brush v. S.F. Newspaper Printing Co., 315 F. Supp. 577, 580 (N.D. Cal. 1970). But an organization may also be treated as an employment agency if it "significantly affects access of any individual to employment opportunities." Scaglione v. Chappaqua Cent. Sch. Dist., 209 F. Supp. 2d 311, 319 (S.D.N.Y. 2002) (quoting Spirt v. Teachers Ins. & Annuity Ass'n, 691 F.2d 1054, 1063 (2d Cir. 1982), *vacated on other grounds sub nom.* Long Island Univ. v. Spirt, 463 U.S. 1223 (1983)). The EEOC has indicated that an organization placing job ads, like a newspaper, counts as an employment agency under Title VII if it "exercise[s] control" over those advertisements or "actively classif[ies] advertisements" as being appropriate for different audiences. EEOC Compl. Man. § 631.2(b)(1) (2009).

While there is little publicly available information about what share of Facebook's advertisement market constitutes job ads, making it difficult to know if job ads are a significant part of Facebook's business, Facebook is a free service that derives its income from advertisements. *See* Gilbert, *supra* note 12. And yet, even if the overall share of Facebook's revenue from job ads is small, when such an ad is placed on Facebook, Facebook exercises nearly total control over which users will see that ad, and who will therefore become aware of that job opportunity. *Id*. Thus, the definitions of what qualifies an organization as an employment agency seem to be, for Facebook, in conflict. If one focuses on the proportion of Facebook's ad sales from job ads, Facebook may not seem like an employment agency under the *Brush* definition. 315 F. Supp at 580. But given the increasing numbers of job ads placed on platforms like Facebook's, Facebook appears to qualify as an employment agency under the *Scaglione* definition. 209 F.Supp.2d at 319.

The fact that Facebook controls ad placement requires resolving this tension in favor of holding Facebook to be an employment agency. Relatively little of its advertising revenue may come from employment ads, but it nonetheless plays a sufficiently significant gatekeeping role as a deliverer of employment advertisements that it ought to be considered an employment agency under Title VII. Indeed, the purposes of Title VII's prohibitions against employment agency discrimination would be undermined

will assume that Facebook and other online advertising platforms are indeed employment agencies under Title VII when they target job ads to users.

Assuming Facebook and similar online advertising platforms are employment agencies, the following Section turns to several challenges that can be brought for employment discrimination. These challenges vary based on the nature of the discrimination alleged.

B. Relevant Title VII Employment Discrimination Actions

Under current legal doctrine, employment discrimination is addressed retroactively through litigation; employers that might otherwise discriminate do not do so out of fear of litigation in response.[101] This Section will describe several kinds of claims that can currently be brought under Title VII in response to a discriminatory employment practice.

Employment discrimination claims fall under two general categories: disparate treatment and disparate impact.[102] Disparate treatment involves intentionally treating people differently on the basis of a protected characteristic,[103] while disparate impact involves a facially neutral practice that nonetheless results in a disparity along such a characteristic.[104] The following Sections will briefly consider each in turn, along with certain subspecifications of each of these doctrines before turning to how these doctrines may be applied to discriminatory online advertisement targeting.

1. Disparate Treatment

Employment discrimination actions brought against employment agencies have thus far all been on the basis of disparate treatment. Disparate treatment occurs when groups are treated differently on the basis of a protected characteristic; it involves discriminatory intent.[105] The Supreme

by finding that an organization that targets and delivers a significant portion of all employment ads is not an employment agency if it is sufficiently large that these job ads make up only a small part of its total business. Thus, as a matter of policy, Facebook should be considered an employment agency under Title VII. Given how prevalent advertising job opportunities online has become, to not find that Facebook is an employment agency is to allow the exception to swallow the rule.

This argument deserves substantially more extensive development and is outside the scope of this Note. It is included here only to show that, although it is not a foregone conclusion that Facebook and similar advertisers would be found to be employment agencies, neither is it unreasonable to argue that they can and should be.

[101] *See infra* note 202 and accompanying text.

[102] JOSEPH G. COOK & JOHN L. SOBIESKI, JR., 4 CIVIL RIGHTS ACTIONS ¶ 21.23 (Matthew Bender & Co. eds., updated 2019), https://advance.lexis.com/api/permalink/37c776b6-725a-4ae0-a96a-073766884322/?context=1000516 [https://perma.cc/N9Q2-HSHK].

[103] *Id.* ¶ 21.22.

[104] *Id.* ¶ 21.23.

[105] *Id.* ¶ 21.22.

Court has held that disparate treatment involves "the refusal to recruit, hire, transfer, or promote [protected] group members on an equal basis with [others]," specifically because they are members of different groups.[106]

Up until Facebook's settlement of various discrimination lawsuits, it required advertisers to specify a gender to which their ads would be displayed ("Male," "Female," or "All").[107] Advertisers were not required to justify this targeting.[108] Facebook was wise to have settled these lawsuits: had the cases proceeded and Facebook been found to be an employment agency, this gender-based targeting would very likely have run afoul of the law, especially since Facebook did not require the advertisers to aver that gender-based targeting represented a BFOQ of employment.[109]

Disparate treatment actions can also be founded on a claim that employment decisions were made in reliance upon stereotypes about protected characteristics.[110] For instance, employers can be held liable for denying a female employee a promotion because she is purportedly insufficiently feminine.[111] Reliance on such stereotypes in employment decisions can form the basis of a claim of disparate treatment and lead to employer liability.[112] One strength of antistereotyping theory is that plaintiffs can state a claim of disparate impact without providing evidence that those who did not share the plaintiff's protected characteristic were treated differently.[113] That is, although a claim for disparate treatment usually involves showing a disparity in treatment between people with the protected characteristic and those without,[114] those same claims brought under an antistereotyping theory only require a showing that stereotypes about the protected group were relied upon in making employment decisions.

2. Disparate Impact

The core of disparate impact doctrine in employment discrimination is that "[Title VII] proscribes not only overt discrimination [i.e., disparate treatment] but also practices that are fair in form, but discriminatory in

---

106 Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 335 (1977).

107 Complaint, *supra* note 15, at 1.

108 *Id*.

109 As noted *supra* note 28 and accompanying text, because Facebook is no longer targeting ads in this way, the problems described in Section I.C of this Note are now the primary areas of concern involving potential discrimination by online advertising platforms like Facebook. *See infra* Part III.

110 Price Waterhouse v. Hopkins, 490 U.S. 228, 251 (1989) (plurality opinion).

111 *Id*. at 235; *see* Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519, 548–49 (2018) (discussing antistereotyping cases).

112 *Price Waterhouse*, 490 U.S. at 251.

113 Bornstein, *supra* note 111, at 549.

114 COOK & SOBIESKI, *supra* note 102, ¶ 21.22.

operation."[115] So if an employment practice appears facially neutral but affects people differently along a protected characteristic, the people negatively affected will have a cause of action under Title VII. The Supreme Court has held that showing discrimination through disparate impact requires not just showing a difference among groups, but also a showing that this difference is not reflected in the pool of qualified applicants.[116] Furthermore, a facially neutral practice that assesses factors related to work qualifications but which may result in a disparate impact, like a written exam, cannot on its own be the basis of a claim of disparate impact.[117]

Courts have used two tests to recognize a disparity in outcomes between groups: statistical significance (performing a statistical analysis to determine whether the disparity is unlikely to be due to chance) and the "four-fifths" rule (checking to see whether one group passes through the process at less than four-fifths the rate of another group).[118] The statistical significance test has the advantage of allowing plaintiffs to use a variety of statistical measures to examine whether disparities are due to chance,[119] some of which may not require precise knowledge of underlying population statistics. The disadvantage is that statistical methods are extremely sensitive to sample size and thus may not be able to detect discrimination on small scales.[120] On the other hand, the four-fifths rule has the advantage of being simple to apply and understand if one has the relevant population statistics,[121] but it sets a fairly high bar for plaintiffs insofar as the disparity must be fairly (even arbitrarily) high to be actionable,[122] which is especially challenging when taking on a fairly large employer or employment agency.

Stating a disparate impact claim requires the plaintiff to point to the specific practice leading to the disparity and to show that a better alternative

---

[115] Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

[116] Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 650–51 (1989); *see also* Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 308 (1977) ("[A] proper comparison [is] between the racial composition of [employees] and the racial composition of the qualified . . . population in the relevant labor market.").

[117] Ricci v. DeStefano, 557 U.S. 557, 579–80, 587 (2009). A facially neutral practice that results in a disparate impact may sustain a claim of disparate impact if it can be shown that a less discriminatory and equally effective alternative exists that the employer refused to use. *Id*.

[118] Jennifer L. Peresie, *Toward a Coherent Test for Disparate Impact Discrimination*, 84 IND. L. J. 773, 774 (2009) (citation omitted). The EEOC uses the four-fifths rule as an acceptable means of demonstrating disparate impact. Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607.4(D) (2019).

[119] Peresie, *supra* note 118, at 785.

[120] *Id.* at 787.

[121] *Id.* at 783.

[122] *Id*. at 782.

practice exists.[123] Both of these requirements have been repeatedly reaffirmed.[124] Finally, even after a prima facie case of disparate impact has been made against a business, that business can raise a defense of business necessity.[125] Whether a business can claim business necessity for a discriminatory practice depends upon "whether a challenged practice serves, in a significant way, the legitimate employment goals of the employer" and upon "the availability of alternative practices to achieve the same [goal], with less [discriminatory] impact."[126] The practice need not be absolutely necessary for the business;[127] instead, it need only advance the business's goals in a way that no nondiscriminatory practice can be shown to do.[128]

One of the ACLU's claims against Facebook in one of the recently-settled lawsuits was that Facebook discriminated by using a practice "legally indistinguishable from word-of-mouth hiring,"[129] which is actionable under a theory of disparate impact.[130] Word-of-mouth hiring occurs when employers advertise jobs to the people they know, and those people in turn tell other people they know about the jobs.[131] For example, in one case, a power company filled employment vacancies by having workers tell their friends about the vacancies; because the workers were mostly white, the people who heard about the job opportunities were overwhelmingly white, causing only a small percentage of the company's labor force to be black.[132]

Word-of-mouth hiring has been held to be discriminatory because it leads to a "circumscribed web of information" about job opportunities,[133] and

---

[123] Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 994 (1988). The *Watson* Court goes on to acknowledge that although identifying the specific practice at fault is "relatively easy to do in challenges to standardized tests, it may sometimes be more difficult when subjective selection criteria are at issue." *Id*.

[124] These requirements were most recently reaffirmed, as of September 2019, in *Texas Department of Housing & Community Affairs v. Inclusive Communities. Project, Inc.*, 135 S. Ct. 2507, 2523 (2015) (reaffirming the requirement that plaintiffs point to the specific practice leading to the disparity); *id*. at 2518 (citing Ricci v. DeStefano, 557 U.S. 557, 578 (2009)) (reaffirming the requirement that plaintiffs provide a better alternative practice).

[125] *Ricci*, 557 U.S. at 578.

[126] Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 658–59 (1989).

[127] *Id*. at 659.

[128] *Id*.

[129] Complaint, *supra* note 15, at 12. The ACLU's claim was that using a lookalike audience to target job ads was equivalent to spreading information about job opportunities by word-of-mouth. *Id.*

[130] United States v. Ga. Power Co., 474 F.2d 906, 925 (5th Cir. 1973) (requiring employer to take affirmative steps to offset the disparate impact of the company's word-of-mouth hiring practices).

[131] *Id*.

[132] *Id*. at 925–26.

[133] *Id*. at 926.

establishes "a *prima facie* case of disparate impact."[134] Unlike other disparate impact claims, word-of-mouth employment discrimination claims are inherently about advertisement, that is, the promulgation of information regarding job opportunities. Nonetheless, there has never been a case where word-of-mouth hiring was the basis of a successful claim against an advertising agency.[135]

III. APPLYING AND ADAPTING EMPLOYMENT DISCRIMINATION TO MACHINE LEARNING

This Note has provided an overview of the basic processes through which employment ads can be targeted to users on social media platforms like Facebook,[136] the mechanics of deep learning neural networks,[137] and the possible employment discrimination actions under current legal doctrine.[138] This Part now begins to integrate these different strands by examining how algorithmic advertising discrimination can be challenged under Title VII, what the obstacles to bringing such challenges are, and how legal scholars have proposed adapting Title VII to the domain of algorithmic employment discrimination.

Despite the challenges posed by DLNN systems as previously described,[139] existing legal doctrine may well be adaptable to these challenges and allow individual plaintiffs redress against employers and employment agencies that discriminate against them. Indeed, Professor Joshua Kroll and his colleagues have suggested that existing antidiscrimination doctrine may not require adaptation: they suggest that any changes to the status quo focus on ensuring algorithms generate sufficient explanations for their decisions rather than adapting employment discrimination to the challenges these algorithms present.[140] They argue that current employment discrimination law may already compel system

---

[134] United States v. Brennan, 650 F.3d 65, 126 (2d Cir. 2011) (citing Grant v. Bethlehem Steel Corp., 635 F.2d 1007, 1016 (2d Cir. 1980)).

[135] A search of WestLaw and Lexis Nexis did not reveal any cases where an advertising agency was successfully sued under a theory of word-of-mouth discrimination.

[136] *Supra* Sections I.A–I.B.

[137] *Supra* Section I.C.

[138] *Supra* Part II.

[139] The main problem with which this Note is concerned is that it can be difficult to prevent a DLNN from learning a discriminatory function, or indeed to discern what function it has learned. *See supra* Section I.C.1.

[140] Kroll et al., *supra* note 29, at 695–705 (2017).

engineers to design their systems for nondiscrimination,[141] which will occur by having engineers design their systems to provide post hoc explanations.[142]

Professor Kroll and his colleagues' proposed solution masks the scale of its ambition and the difficulty of its realization by shifting it outside the context of discrimination law. Having systems like DLNNs generate reliable post hoc explanations is an open problem in artificial intelligence: requiring engineers to have their systems be able to generate explanations may result in such systems being made functionally illegal for many purposes, which no scholars have endorsed.[143] Moreover, because systems for explaining DLNN decision-making are not yet universally reliable, clear, and consistent,[144] encouraging or even requiring engineers to have their systems provide the best explanation they can generate may not even be helpful in a given case if the explanation does not clearly resolve the issues.

Nonetheless, Professor Kroll and colleagues are correct that if the facts underlying employment discrimination can be discovered, existing employment discrimination actions may well suffice to provide relief for individual plaintiffs. Whether such facts are discoverable, however, has not been addressed. This Note now turns to the actions available to such plaintiffs, specifically examining whether the facts required to state a claim

---

[141] *Id*. at 694–95. It is not clear whether Professor Kroll and his colleagues believe that current employment discrimination law legally *compels* designing for nondiscrimination or if they see this question as irrelevant. Regardless, they focus on technologically driven, rather than legally driven, solutions, arguing that the main responsibilities that lawmakers and policymakers have are to clarify what the law is and to maintain vigilance over and literacy with regard to how such systems are used. *Id*. at 699–704.

[142] *Id*. at 698. Kroll and colleagues acknowledge that it is not entirely clear how Title VII might be applied to a discriminatory algorithm and that it is possible that claims against discriminatory algorithms might be rejected under current legal doctrine, *id.* at 693–95, but they ultimately conclude that the solution to this problem lies in changing the algorithms rather than legal doctrine, *id.* at 696–99. *See also* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV 671, 729–32 (2016) (arguing that the problem lies not with Title VII, but with the explanations generated by data mining in ML systems: where machine learning systems introduce new biases and forms of discrimination into the workplace, they argue that Title VII should be sufficient to address and rectify the problem, but where ML systems instead replicate and propagate disparities that exist in broader society, Title VII is the wrong tool for the job, and the designers of the systems should not be held responsible); Kleinberg et al., *supra* note 55, at 51 (arguing that, if steps are taken to ensure ML systems are developed transparently, "[t]he use of algorithms offers far greater clarity and transparency about the ingredients and motivations of decisions, and hence far greater opportunity to ferret out discrimination").

[143] The explanation systems described *supra* Section I.C.2 hold promise but are not universally and consistently reliable across ML systems. If the law requires systems to provide explanations but the technology cannot reliably generate such explanations, the systems are in violation of the law until the explanation systems improve. In any event, having the ability to generate such explanations would not necessarily mean the problem was solved.

[144] *See supra* Section I.C.2.

can be discovered under existing doctrine and discussing changes to employment discrimination law that scholars have proposed to make these facts more discoverable.

A. Disparate Treatment

Disparate treatment occurs when groups are treated differently on the basis of a protected characteristic and involves discriminatory intent.[145] Disparate treatment might have been a legally adequate cause of action as long as Facebook allowed targeting job ads explicitly on the basis of gender or of variables that are clearly proxies for gender.[146] But in light of Facebook's settlement prohibiting such explicit targeting,[147] plaintiffs will likely not be able to make out a claim of disparate treatment against Facebook's advertising placement practices. Indeed, once protected characteristics are removed from the data upon which machine learning systems train and operate, it will be nearly impossible to demonstrate that those systems are treating people differently on the basis of those characteristics—even if they are.[148] If the data with which the algorithm are trained encode the biases of the people who collected the data, the algorithm will likely be biased. But there may be no single actor or institution at any point in the process who demonstrably displayed a biased intent, especially if the system was trained on a dataset compiled by a third party.[149]

A plaintiff might be able to show that the training data were generated or manipulated in such a way that a court could infer that the actor's intention was to train a discriminatory system.[150] But it also is possible that no humans

---

[145] *See supra* Section II.B.1.

[146] Such claims may still prove workable against other ad platforms that allow such targeting. Nonetheless, Professor Kroll and his colleagues provide an insightful warning against bringing such claims. An allegation of disparate treatment based only on the fact that

> the design of the algorithm includes inputs that are a proxy for class membership . . . would be valid against virtually any system with a significant number of inputs. It seems more likely that courts would reject the formal-rule subset of disparate treatment for algorithmic decisions than that they would hold the majority of algorithmic decision-making to constitute disparate treatment.

Kroll et al., *supra* note 29, at 695.

[147] *See supra* note 16 and accompanying text.

[148] Such information may be latently encoded in the rest of the data. *See supra* Section I.C.1.

[149] Given that many separate actors are often involved in designing what features a dataset should encode, assembling that dataset, designing a DLNN's architecture, and actually training a DLNN system on that dataset, assigning blame for the DLNN's actions is a challenging problem and is outside of the scope of this Note.

[150] Barocas & Selbst, *supra* note 142, at 692–93; Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 884–85 (2017); Kroll et al., *supra* note 29, at 681–82.

involved in collecting data or training a system will have behaved with discriminatory intent or outright treated groups differently on the basis of a protected characteristic, but that the algorithm that results from the interplay of their actions does. The doctrine of disparate treatment will either have to adapt to these facts or will not be the avenue through which algorithmic discrimination will be challenged.

B. Antistereotyping Theory

Professor Stephanie Bornstein has proposed that algorithmic discrimination can be challenged in court as being based in and reflective of stereotypes.[151] She incisively notes that the fact that antistereotyping theory does not require comparing the plaintiff's treatment against the treatment of others is an enormous benefit in the domain of algorithmic discrimination.[152] In the context of advertising discrimination in particular, where it may be extremely difficult to uncover information regarding the treatment of the underlying base group, this is a significant advantage for plaintiffs.[153] And indeed, as discussed in Part I, machine learning algorithms may well be learning something akin to a stereotype, that is, a function that encodes that people of *this* category tend to result in *that* outcome.[154] The problem with using an antistereotyping theory is that there is no way to show that this is, in fact, the function that the algorithm has learned.[155]

Antistereotype claims brought under Title VII are claims of disparate treatment.[156] As the Court in *Price Waterhouse* explained, "[t]he plaintiff must show that the employer *actually relied on* [the forbidden characteristic] in making its decision."[157] In a plaintiff's best-case scenario, she would have to show that "the algorithm is trained on data that itself incorporates subjective biases" and is replicating them;[158] at worst, she would have to

---

[151] Bornstein, *supra* note 111, at 549–50.

[152] *Id*. at 549.

[153] It is difficult to uncover such information in this context because the ads are shown in real time on the private profiles of individual users. Unless all such ads were to be monitored, it may be impossible for a plaintiff to determine who saw which ads. Such information may well be obtainable through discovery, but the plaintiff must state a valid claim before getting to discovery, Ashcroft v. Iqbal, 556 U.S. 662, 678–79 (2009). It is unlikely that a company like Facebook would volunteer to potential plaintiffs the information necessary to state a claim against it. Thus the information required to get to discovery may only be accessible to plaintiffs through discovery.

[154] *See supra* notes 52–62 and accompanying text.

[155] An antistereotyping theory may be effective against decisions by more interpretable ML systems, like nearest neighbor. *See generally* Cover & Hart, *supra* note 43.

[156] Price Waterhouse v. Hopkins, 490 U.S. 228, 251 (1989) (plurality opinion).

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

[158] Bornstein, *supra* note 111, at 562.

show that the algorithm had extracted harmful *latent* stereotypes from the dataset upon which it was trained. In the first case, relying upon an antistereotyping theory would therefore only move the goalposts from showing that the algorithmic decision-maker relied upon stereotypes to showing that the people who generated the training data relied upon them. In the second, the difficulties involved in showing what the system had learned would directly translate into a difficulty in showing it was making its predictions on a forbidden basis.[159]

C. Disparate Impact

If employment ads on online platforms are not being targeted explicitly on the basis of a protected characteristic, then they will be illegally discriminatory if they are shown to one group much more than another; that is, if their targeting results in a disparate impact among groups.[160] But once Facebook removes protected characteristics from the bases for advertising, this discrimination will become apparent only by examining the entire population of people to whom the ad is shown. Who saw which ad is likely information that can be discerned through discovery, but plaintiffs still must successfully state a plausible claim for relief before they can proceed to discovery.[161] Additionally, showing that the group of people who see an ad does not reflect the underlying pool of candidates may require accessing the data of Facebook users generally, which raises serious privacy issues.[162] And showing that there is an available nondiscriminatory alternative,[163] such as using bias mitigating approaches, would require inspecting the actual algorithms used by companies like Facebook—algorithms that form the basis of their revenue-raising business and are fiercely guarded trade secrets.[164]

---

[159] Kroll and his colleagues' warning holds true in this area about how plaintiffs' overclaiming the potential for algorithmic discrimination could hurt efforts to bring such claims. *See supra* note 146 and accompanying text. A claim that a DLNN's *potential* to be discriminating on the basis of stereotypes might lead courts to broadly reject such challenges as potentially attaching too much liability to too many actors.

[160] *See supra* Section II.B.2.

[161] Ashcroft v. Iqbal, 556 U.S. 662, 678–79 (2009); *see supra* note 153 and accompanying text.

[162] *But see* Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 124–27 (2014) (suggesting a procedure for due process that includes a hearing with a trusted third party "act[ing] as a neutral data arbiter to routinely examine Big Data providers").

[163] *See supra* notes 123–124 and accompanying text.

[164] *See, e.g.*, David S. Levine, *Confidentiality Creep and Opportunistic Privacy*, 20 TUL. J. TECH. & INTELL. PROP. 11, 28–29 (2017). Additionally, explaining the complex, technical nature of these algorithms will be an obstacle to litigators, although hardly an insurmountable one.

Even if disparate impact is shown, it may be nearly impossible for a plaintiff to narrow down the specific *cause* of that disparate impact.[165] Assigning blame to either advertisers expressing their preferences, biases and dataset artifacts amongst the algorithm's training data, or the company that actually trained the algorithm, may be extremely difficult. Yet recovering against an employment agency requires finding the employment agency responsible for the discrimination.[166] Thus, stating a claim against a company like Facebook for how its algorithm placed an ad is complicated: no one, including the people who built it, will know exactly what the algorithm is doing, except by reference to the observable actions it takes.[167]

Professor Pauline Kim has argued that Title VII could be read to directly prohibit disparate outcomes stemming from the use of machine learning systems, without necessarily requiring plaintiffs to identify the specific practice used nor to show that a better practice is available.[168] However, she recognizes that even if plaintiffs are only required to show a disparity in outcomes and nothing more, discovering such a disparity can be enormously challenging because the data that encode those disparities are privately held by companies.[169] Therefore, Professor Kim argues that discovery rules should be modified in the employment context to allow plaintiffs to better gather the information necessary to state a claim for disparate impact resulting from the use of ML systems.[170]

Professor Kim's solution seems sound and particularly workable with regard to the domain of hiring and promotion decisions. When companies hire or promote people, there is a circumscribed universe from which potential candidates are selected (i.e., applicants to a job or those eligible for promotion), so it is relatively straightforward to compare the group that *was* advanced against the entire selection pool. But it may not catch such discrimination in the domain of employment *advertising*. For advertising, the comparison group is substantially more nebulous, comprising all people

---

[165] Kim, *supra* note 150, at 909 ("[T]he *concept* of disparate impact . . . applies to classification bias. The problem is that the ways the doctrine has been applied in the past are not well suited to address the data-driven nature of classification bias.").

[166] 42 U.S.C. § 2000e-2 (2012).

[167] *See supra* Section I.C.1. As discussed *supra* Section I.C.2, there is a certain extent to which individual decisions might be explainable, but explaining the workings of the system as a whole is a significant technical, to say nothing of legal, challenge.

[168] Kim, *supra* note 150, at 910–12.

[169] *Id.* at 919.

[170] *Id*. at 917–20*; see also* King & Mrkonich, *supra* note 54, at 567 (explaining how plaintiffs in such actions will have to obtain information crucial to their claim through discovery).

who possibly could have been shown an ad.[171] Depending on the nature of the advertisement, this could conceivably include all users of a social-media system.

Furthermore, companies such as Facebook may be able to reasonably make claims of business necessity, even if it is possible to show that their ad-placing systems are discriminatory.[172] Facebook (and Google, and other such platforms) make the bulk of their revenue by targeting ads to users on the basis of those users' profiles,[173] and they must rely on ML algorithms to do so because targeting ads by hand is impracticable with billions of users and hundreds of thousands, if not millions, of advertisers. And there is no guarantee that humans would do a "better"—that is, a less biased—job. These companies can therefore make a compelling argument that biased or not, they have no choice but to use these systems.[174]

Yet arguing that ML algorithms are necessary and unavoidable misses the point. The argument is not that ML algorithms should not be used but that the engineers must do everything they can to correct the biases within those algorithms. But such changes involve costs. There will generally be an argument that taking any steps to ensure those algorithms are as unbiased as possible involves high engineering and computational costs and no guarantees of success, even for companies with ample resources, like Facebook.[175] Under current legal doctrine, this argument is likely to be found

---

[171] The comparison group who did not see an advertisement will be ill-defined in any advertising discrimination claim, not just employment discrimination.

[172] *See supra* Section II.B.2.

[173] Gilbert, *supra* note 12.

[174] Similarly, Professor Kim as well as Professors Barocas and Selbst have noted that systems that learn who to hire and promote based on past hiring and promotion data would pass a requirement that they be related to successful prediction as they are by definition learning a predictive model. *See* Kim, *supra* note 150, at 866; Barocas & Selbst, *supra* note 142, at 708–09. ML systems that predict past patterns are, in fact, predicting those patterns. Professor Kim points out that this is a compelling argument for *not* treating ML systems as unbiased simply because they are good at predicting past outcomes. Kim, *supra* note 150, at 866.

Professor Kim's point has merit in that regard, but hiring and promotion systems are fundamentally different from ML ad-placement systems: ad-placement systems do not operate over a circumscribed set of possible hires. *See supra* note 171 and accompanying text. Such systems have less of a claim to be related to job performance than a system that predicts hiring decisions: unlike hiring or promotion decisions, where the system clearly know who was *not* hired or promoted, advertising decisions derive directly from the pool of people being targeted, without a clear comparison group to learn the salient features that meaningfully differentiate them from the general population and which can predict actual job performance, *see supra* Section I.B. The system, therefore, will not learn to drop useless, nonpredictive characteristics. There is no way to know whether someone would have clicked on a job ad and would be qualified for that job if they were not even shown the ad. Thus, ad-placement systems have a weaker claim to be related to eventual job performance than hiring and promoting systems.

[175] *See supra* Section I.C.2.

persuasive by default,[176] which could allow companies like Facebook to avoid taking *any* such steps. At a minimum this must change, so that such companies must take reasonable steps towards unbiasing their systems.[177]

Automated systems that learn their own decision-making criteria present a fundamentally new kind of actor in the employment discrimination context.[178] Up until now, the legal system has dealt with inert tools like questionnaires and tests—which, so long as they are facially neutral and a better option is not available, are permitted to lead to disparate impacts upon the populations those tools assess[179]—and humans, who are not allowed to make decisions on the basis of forbidden characteristics. Machine learning systems are capable of being facially neutral, like a test, while learning to make decisions using characteristics upon which humans would not be allowed to base decisions.[180] These systems therefore may be too neutral to be caught by a theory of disparate treatment and not biased enough to be prohibited under a theory of disparate impact.

D. Word-of-Mouth Hiring

Because Facebook settled the ACLU's lawsuit against it, it remains unclear whether the ACLU's challenge against Facebook's lookalike audience tool under principles of word-of-mouth hiring would have been successful. Word-of-mouth hiring certainly has parallels to what Facebook does. Word-of-mouth hiring is not allowed because of the assumption that the people passing along job openings know other people like them, and advertising only to people that "look like" the people an individual knows "circumscribe[s the] web of information" in violation of Title VII.[181] Similarly, advertising to friends of people already being shown ads is only one step removed from traditional word-of-mouth hiring.

But the advertising techniques used by Facebook and similar online advertising platforms are not analogous to word-of-mouth hiring in key ways. Advertisers can target users quite different and disconnected from

---

[176] *See supra* notes 125–128 and accompanying text.

[177] Kroll and his colleagues agree that encouraging developers to do everything possible to design systems not to discriminate is desirable. Kroll, *supra* note 29, at 695. But they argue that system engineers simply *should* design their systems not to discriminate, even if Title VII does not outright require them to do so. *Id.* at 694–95. Although that would be an ideal scenario, it seems more likely that corporations would avoid voluntarily making the necessary, costly development of their algorithms.

[178] *See id.* at 693.

[179] Provided they assess things related to work qualifications. Ricci v. DeStefano, 557 U.S. 572, 579 (2009).

[180] *See supra* Section I.C.1.

[181] United States v. Ga. Power Co., 474 F.2d 906, 925–26 (1973).

their base group, suggesting the "web of information" is not circumscribed to people highly similar to members of the original group.[182] If the lookalike audience does not sufficiently resemble the advertiser-supplied audience, a court might find that the reasons to forbid word-of-mouth hiring are not implicated. And generating advertising audiences using lookalike technology is literally not word-of-mouth hiring: the people seeing the ads may have absolutely no social connection to the base group. Thus, if a court strictly construes principles of word-of-mouth hiring prohibitions, tools like Facebook's lookalike audience tool will likely not be found to be a word-of-mouth hiring practice.

E. Reckless Discrimination

Professor Bornstein has additionally proposed a new theory of liability for discrimination caused through recklessness, one that could address the shortcomings in disparate impact liability actions.[183] Aiming to address the problem of employers failing to make any effort to correct for the problem of implicit bias, Professor Bornstein proposes that liability under Title VII should accrue to employers who ignore well-documented risks of bias and do not take well-established steps to reduce those risks.[184] Under this theory, a conscious disregard for a well-established risk that results in that risk coming to fruition should be found to be sufficiently similar to intending the undesirable outcome that the employer is held liable under Title VII.[185] This theory of liability would be based on disparate treatment (with reckless intent) rather than pure disparate impact.[186]

The domain of implicit bias in humans is somewhat analogous to the problem of algorithmic bias in DLNN systems. Implicit bias in humans involves the unconscious activation of an attitude or stereotype about a person based on their group membership.[187] Based on this activation, human decisions and judgments about a person can be affected by the knowledge that that person is a member of that particular group.[188] Similarly, the concern

---

[182] "Creating a larger audience increases your potential reach, but reduces the level of similarity between the Lookalike Audience and source audience." *About Lookalike Audiences*, FACEBOOK: BUSINESS, https://www.facebook.com/business/help/164749007013531 [https://perma.cc/F9GU-9LLY].

[183] Stephanie Bornstein, *Reckless Discrimination*, 105 CALIF. L. REV. 1055, 1055–56 (2017).

[184] *Id.* at 1103.

[185] *Id.*

[186] *Id.* at 1105.

[187] *See, e.g.*, Anthony G. Greenwald & Mahzarin R. Banaji, *Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes*, 102 PSYCH. REV. 4, 7–10, 14–18 (1995).

[188] *Id.*

with DLNNs is that latent patterns in an individual's data will lead the system to discern and rely upon that individual's membership in a protected group.[189]

People are by definition not aware when their decisions involve unconscious biases; when they attend to the source of the bias, the bias is reduced.[190] So once they are aware of the potential for bias, employers and individuals can take effective steps to reduce its impact.[191] Similarly, ML researchers are sufficiently aware of the problem of algorithmic bias that researchers are working on ways to mitigate the problem,[192] and Facebook has now agreed to police at least the potential biasing use of information that acts as a proxy for protected characteristics.[193] Professor Bornstein makes a compelling argument that, given the well-documented prevalence of implicit biases, employers should almost be required to take such preventative steps.[194]

Adopting a doctrine of reckless discrimination under Title VII, particularly in the domain of algorithmic decision-making, could alleviate some of the burdens plaintiffs must bear when they seek to connect an employer's use of an algorithm to some negative outcome. Because this theory is grounded in disparate treatment rather than disparate impact, it would reduce the availability of the business necessity defense.[195] It would also place the responsibility squarely on the employer rather than the system the employer relies upon. As a result, employers would be incentivized to take every countermeasure possible to combat algorithmic discrimination, and the theory would not let them off the hook if they understood that, despite those countermeasures, there was still a substantial risk of the algorithm they used being discriminatory. There would still be obstacles to plaintiffs discerning the information required to bring a claim, but supplementing this theory with Professor Kim's proposals to make discovery more accessible[196] could enable plaintiffs to hold parties whose algorithms discriminate against them responsible.

The foregoing analysis demonstrates that employment discrimination law is not toothless in the face of algorithmic advertising discrimination, provided that corporations like Facebook are found to be employment

---

189 *See supra* Section I.C.1.

190 Greenwald & Banaji, *supra* note 187, at 18.

191 Bornstein, *supra* note 183, at 1096.

192 *See, e.g.*, Zhang et al., *supra* note 54, at 335.

193 Settlement, *supra* note 16, at 1.

194 Bornstein, *supra* note 183, at 1103–07.

195 *See supra* Section II.B.

196 *See supra* note 170 and accompanying text.

agencies for purposes of Title VII when they target job ads to individual users. Professor Kroll and his colleagues show how requiring the users of algorithms to generate explanations for their algorithms' decisions can provide sufficient information to allow Title VII plaintiffs to bring actions to counter employment discrimination.[197] Professor Kim, in turn, explains how disparate impact actions can adapt to the challenges posed by DLNNs by loosening discovery requirements and allowing plaintiffs to state claims based on pure disparate impact without allowing the defense that the algorithm is making decisions reasonably related to future employment performance.[198] And Professor Bornstein's theory of reckless discrimination would allow courts to hold employers and employment agencies like Facebook liable for their discriminatory algorithms, despite any claims of business necessity, if those businesses have not taken all steps reasonably possible to mitigate the possibility of such discrimination.[199]

IV. COUNTERING ALGORITHMIC DISCRIMINATION REACTIVELY AND PREVENTATIVELY

The question remains whether the adaptations described in the previous Part, and the threat of litigation under such claims, suffice to address algorithmic employment discrimination. If the goal of employment discrimination law is to give recourse to aggrieved parties, then these adaptations may be enough. But if employment discrimination law exists to prevent employment discrimination from occurring,[200] it remains to be seen whether litigation will achieve this goal.

This Part concludes that, although these adaptations may provide relief to individual plaintiffs and even classes of plaintiffs, they are insufficient to adequately prevent such discrimination from occurring in the first place.[201] Litigation is fundamentally a reactive solution, which prevents undesirable behavior through deterrence.[202] But because even maximally deterred ML developers may be unable to prevent their systems from discriminating,

---

[197] Kroll et al., *supra* note 29, at 696–99.

[198] Kim, *supra* note 150, at 917–20.

[199] Bornstein, *supra* note 183, at 1103.

[200] The fact that Title VII defines a wide range of discriminatory employment behaviors as unlawful suggests that the law is designed to prevent such discrimination, not only provide redress for it. *See* 42 U.S.C. § 2000e-2 (2012).

[201] *See infra* Section IV.A.

[202] *See* STEVEN SHAVELL, FOUNDATIONS OF ECONOMIC ANALYSIS OF THE LAW 93 (2004) ("[F]inancial incentives [can] reduce harmful externalities. Under a *liability rule*, parties who suffer harm can bring suit against injurers and obtain compensation for their losses, motivating injurers to avoid causing harm.").

preventing algorithmic discrimination by DLNNs requires a proactive solution.[203] Specifically, the most effective way to prevent an ML system from learning the wrong thing is to train it on a sufficiently large and diverse dataset such that random biasing patterns are less likely to appear in that dataset.[204] In order to use techniques like bias mitigation, this dataset must contain extensive private information about individuals and their protected characteristics.

This Note proposes a proactive solution, allowing ML engineers to train their systems on a large, rich, diverse dataset of real people's data.[205] Because the existence of such a dataset poses serious privacy concerns for the people whose data are included in it, this data should not be publicly available, nor should it simply be given to engineers training ML systems. Instead, some specialized third party, such as a nonprofit organization, an industry consortium, a public–private venture, or a new government agency, should be tasked with maintaining and safeguarding this dataset and with testing systems against this dataset for bias.

A. The Insufficiency of Reactive Solutions

Although reckless discrimination provides the most promising path forward for challenging algorithmic discrimination under Title VII, all the theories discussed above are valuable contributions which, if followed, would almost certainly make it easier to challenge algorithmic discrimination in the courts.[206] Courts should look to these proposed solutions as ways to address antidiscrimination litigation currently moving through the legal system and the cases that will be brought in the near future.[207]

But ultimately, the question remains whether adaptations that focus on how to recognize and challenge algorithmic discrimination in court would only treat the symptoms of algorithmic bias rather than seek to cure it. A credible threat of successful litigation may well be enough to induce DLNN engineers to implement techniques like bias mitigation, or other techniques that might be developed to reduce the potential for discriminatory bias. But

---

[203] *See infra* Section IV.A. Kroll and colleagues agree that designing for antidiscrimination must be a component of algorithmic development but conclude that legal mechanisms beyond the threat of litigation are unnecessary. *See* Kroll et al., *supra* note 29, at 694–95.

[204] *See supra* Section I.C.1.

[205] *See infra* Section IV.B.2.

[206] *See supra* Part III.

[207] *See, e.g.*, *supra* note 16 and accompanying text (discussing HUD's recently filed action against Facebook).

the discussion in Part I of this Note presents a hard truth: no matter the intentions of a DLNN's engineers or the steps taken to prevent such systems from discriminating, they will not know what their system is doing until it is deployed on extensive real-world data.[208] Because of this, once the DLNN system has been trained and deployed, it can be too late to fix the problems it may create.[209]

Thus, the threat of litigation is insufficient to fix the problem. Litigation may incentivize engineers to take every step possible to mitigate the risk of discrimination, but these may not be enough to actually avoid that discrimination if their systems are trained on data insufficiently rich to allow their systems to learn a useful function without learning a discriminatory one.[210] It is unrealistic to demand that every company build its own sufficiently large, reliable datasets, as doing so involves enormous human capital costs.[211]

Furthermore, when a plaintiff prevails in her case, it is not clear what should happen next. Will the system that had been found to discriminate against that particular plaintiff be allowed to continue to be used on others? Should the company that created the system be required to scrap it entirely and start over, potentially at enormous cost to their business? Both options have obvious shortcomings. And unless the company was prevented from using DLNNs at all, there would be no mechanism to ensure that whatever system would subsequently be put into use would not discriminate against other people as well.

Forbidding companies from using DLNNs in employment contexts, or other contexts in which discrimination is a concern, would be a repudiation of progress and technological development and likely significant overkill. Instead, if we want to go beyond giving recourse to individual plaintiffs who have the time and resources to pursue a claim against corporations like Facebook and to develop a legal mechanism designed to prevent such

---

[208] *See supra* Section I.C.1.

[209] Even if the system can be pulled before it does any harm, the system's engineers would have to retrain and redeploy the system on whatever new data they gathered—an expensive and time-consuming process, with no guarantee that the updated system would be significantly better or would not have to be immediately pulled again. Demanding that businesses repeatedly shoulder these costs is unrealistic and unfair. *See supra* Section I.C.1; *see also* Turney, *supra* note 81, at 4–5.

[210] *See supra* Section I.C.

[211] *See* Turney, *supra* note 81, at 3; Gary M. Weiss & Foster Provost, *Learning when Training Data Are Costly: The Effect of Class Distribution on Tree Induction*, 19 J. ARTIFICIAL INTELLIGENCE RES. 315, 315 (2003). Requiring companies to build their own datasets or go find one put together by someone else increases the risk of poisoning attacks. *See supra* note 68 and accompanying text.

discrimination from occurring in the first place, those legal mechanisms must be proactive rather than reactive.[212]

B. Proactive Solutions

This Section examines recently proposed legislation designed to vet ML systems before their deployment. It then proposes a new path forward for countering algorithmic discrimination: an organization charged first with creating and maintaining a dataset sufficiently representative of the general population that ML systems trained upon it are less likely to be discriminatory, and second with monitoring and verifying the performance of those same ML systems.

1. The Algorithmic Accountability Act

Legislation was recently introduced in Congress to create a mechanism to vet artificial intelligence systems prior to their being rolled out to customers (and to retroactively vet already-deployed systems).[213] This Bill, the Algorithmic Accountability Act (the Act), would give the Federal Trade Commission (FTC) authority to issue rules and regulations regarding the development of certain artificial intelligence systems, set standards for unacceptable levels of unfairness, discrimination, or data vulnerability in those systems, and provide some oversight over system developers.[214] The Act does not define what would count as unfair or discriminatory outcomes caused by the regulated systems, nor what would be required to remediate them, leaving those determinations to the FTC in its rulemaking capacity.[215] But the Act would, at a minimum, require certain entities—those that produce certain kinds of artificial intelligence systems—to also produce

---

[212] The reader may think that this Note is simply restating the proposed solution of Kroll et al., *see supra* notes 140–142 and accompanying text. Not so. It certainly agrees with Kroll and his colleagues that the solution to the problem of discriminatory algorithms requires the algorithms to be built properly in the first place. But whereas Kroll and his colleagues appear to argue that system engineers may well take all appropriate steps themselves, Kroll et al., *supra* note 29 at 694–95, this Note argues that they must be *compelled* to do so. Kroll and his colleagues also appear to believe that Title VII litigation requires no adaptation for plaintiffs aggrieved by algorithms, *id.*, a position with which this Note disagrees. Most importantly, Kroll and his colleagues do not address the fact that to build the kinds of antidiscriminatory ML systems that they—and this Note—want engineers to build, those engineers need access to a larger and richer database than they are likely to have access to. It is the need for access to such a database that leads this Note to conclude that some third-party entity is needed. *See infra* Section IV.B.2.

[213] Algorithmic Accountability Act, S. 1108, 116th Cong. (2019).

[214] *Id.* § 3(b)(1).

[215] *Id.*

reports to the FTC.[216] Those reports are to assess the systems' risks and benefits, and to take whatever remedial steps the FTC deems necessary to remedy the problems the reports identify.[217]

The Act is a good start, aimed as it is at catching and preventing the harmful outcomes that such systems might create before they come to pass.[218] But ultimately, although the Act could significantly improve upon the status quo, it appears insufficient to consistently prevent the discriminatory outcomes described herein from occurring. As an initial matter, the Act only covers entities that make more than $50 million in annual revenue, have more than a million users' personal information, or act as data brokers.[219] By defining covered entities by what they have rather than by what they do, the Act would permit a small company to develop a discriminatory system with wide-reaching effects so long as the company maintains less than $50 million in revenue, trains its system on fewer than one million users' data, and doesn't sell that data to others. Furthermore, the Act requires those entities to assess and report on their own system rather than having outsiders evaluate them and does not mandate that the systems be tested on new data.[220] As described above, developers may be unaware that their systems discriminate until those systems are deployed and operate upon new data, limiting the value and effectiveness of such self-reporting.[221]

And therein lies the rub. Although the Act, by establishing a proactive assessment system, improves upon the purely reactive options, it leaves the FTC with the same limited options a court would have when confronted with a discriminatory system whose developers took every step possible to avoid having their system discriminate.[222] Like a court, the FTC could either allow that system to be deployed in full or block it entirely; there is no mechanism to support the developer's efforts to make a nondiscriminatory system in the first place. And once again, if a system's developer has a limited or biased

---

[216] *Id.* § 2(5) (covered entities include those with over $50 million in annual revenue, those with over one million users' personal information, and data brokers).

[217] *Id.* §§ 2(2), 2(5), 3(b)(1). The Act is extremely thorough in its definition of the risks such systems can pose. *Id.* §§ 2(7)–(8).

[218] Of course, it is difficult to know how effective the Act would be until the FTC released regulations giving it effect; if the FTC decides that addressing an identified problem involves little more than an ineffective disclaimer, the Act would have little effect.

[219] S. 1108 § 2(5).

[220] *Id.* § 3(b)(1). The Act does "require each covered entity to conduct the impact assessments . . . , if reasonably possible, in consultation with external third parties . . . ." *Id.* § 3(b)(1)(C). It is unclear how "if reasonably possible" would interact with, for example, trade secret laws.

[221] *See supra* Section I.C.

[222] *Supra* Section I.A.

dataset with which to train the algorithm, there may be nothing that developer can do to avoid creating a biased algorithm.[223]

2. The Right Dataset in the Right Hands

If ML systems are to be prevented from discriminating, such systems must be trained on enormous and representative datasets that, to the greatest extent possible, do not encode forbidden biases that the systems can pick up on.[224] These datasets must include variables upon which it is forbidden for systems to discriminate (such as sex, race, religion, etc.)—variables which the engineers of these systems may well not have access to, and which the end-users of the systems may be extremely loath to provide to those engineers.[225] These datasets must contain real human data since artificial data is unlikely to replicate the many patterns latent therein.[226] And these systems will have to be tested on yet another enormous, representative dataset, to ensure that the system did not, in fact, learn a discriminatory algorithm, nor did it learn a function to "finesse" the particular dataset it trained upon—the equivalent of "teaching for the test."[227] If the creators of these systems have access to this latter dataset as they train their systems, the purposes of keeping the testing dataset separate will be undermined as the creators may be able to use this testing dataset during training.[228] Thus, regardless of where developers get their training data, the test data must be managed by some third party.

The only feasible, long-term solution to these challenges is the creation of an external organization—which, for clarity, this Note will call "the Recordkeeper"—that can manage both a training and a testing database[229] of real human data and make them available to ML engineers for training and testing, while not actually exposing those engineers to the private information contained therein. Because the data must be representative of the entire underlying population and will include extensive private

---

[223] *Supra* Section I.C.1.

[224] *See supra* Section I.C.1. One source of algorithmic bias is the plethora of functions that can explain the data. The larger and richer the dataset, the fewer such functions there will be.

[225] *See supra* Section I.C.2.

[226] *See, e.g.*, John Murray, *Training AI with Fake Data: A Flawed Solution?*, BINARY DISTRICT (Oct. 8, 2018), https://journal.binarydistrict.com/can-you-spot-a-fake-training-machine-learning-algorithms-with-synthetic-data [https://perma.cc/KFV7-H37V].

[227] *See supra* Section I.C.1.

[228] Domingos, *supra* note 51, at 80.

[229] In practice, there would likely be a single database that could be carved into different training and testing portions each time such datasets were needed. *See id.*

information, the best way to gather it may be simply to pay people a small amount to voluntarily participate and provide their data.[230]

The Recordkeeper ought to make this data available for system developers to train on, not just to test on, for several practical considerations. First, in order to avoid building discriminatory systems, developers will need access to such a dataset anyways. Assembling it will be enormously time-, resource-, and labor-intensive, so it makes sense to have all interested parties share in those costs rather than to have each redundantly bear them. Although a company like Facebook might be able to bear such costs, less wealthy actors might be prevented from entering the marketplace if they cannot afford to assemble their own. Furthermore, by definition a single, highly secured repository of information is more secure than multiple copies held by many different actors, which is only as secure as its least secure copy.

Finally, these databases will include private information about the people represented in them, including information regarding protected characteristics, both to enable bias mitigation training and to detect bias when it manifests. Because of the sensitivity of this information, access to these databases should be as restricted as possible; the datasets should serve as black-box training grounds that Facebook and others can access only for training and testing purposes, without ever accessing the information itself. The Recordkeeper itself should not be able to examine the data beyond what is necessary to build and maintain the system, nor should it be permitted to share the data with anyone else, including the government.[231] Instead, these datasets could be made available to system engineers one datum at a time for use in training and testing, encrypted such that the data could not be copied by those engineers. Or the Recordkeeper could be charged with developing

---

[230] Any law mandating that individuals *must* share the highly sensitive personal data discussed herein might be found unconstitutional under the Fourth Amendment because people have a reasonable expectation of privacy in such data. *See, e.g.*, Riley v. California, 134 S. Ct. 2473, 2489–90 (2014) (holding that a person's cell phone cannot be searched without a warrant partly because it can contain "[t]he sum of an individual's private life"). Such data could conceivably be collected through use of administrative warrants. *See* Camara v. Mun. Court of City & Cty. of S.F., 387 U.S. 523, 538 (1967) (holding that, in certain administrative contexts, "'probable cause' to issue a warrant to [search] must exist if reasonable legislative or administrative standards for conducting [that search] are satisfied"). Significant political opposition could be expected if the government sought to mandate disclosure of such private data. *See, e.g.*, Michael Price & Faiza Patel, *Muslim Registry or NSEERS Reboot Would Be Unconstitutional*, LAWFARE (Nov. 22, 2016, 12:45 PM), https://www.lawfareblog.com/muslim-registry-or-nseers-reboot-would-be-unconstitutional [https://perma.cc/X336-CYEF]. This Note therefore instead proposes getting people to opt-in to such a system with financial incentives and by motivating them to be a part of combating algorithmic discrimination.

[231] The data should be used only for the purposes described herein because people may be unlikely to volunteer such information about themselves if they think it will be shared, either with third parties or the government.

training environments that allowed the data to be transferred to engineers' systems for training, but in a sufficiently controlled fashion that siphoning off of or access to the data for anything other than training purposes could be detected.[232] By making a dedicated Recordkeeper responsible for collecting and managing the database, companies would never need to solicit information on protected characteristics from their customers just to ensure they are not discriminating against them.[233]

In addition to managing these databases, the Recordkeeper would be responsible for ensuring that system engineers use all appropriate state-of-the-art techniques to avoid training discriminatory systems, such as bias mitigation. And the Recordkeeper would be charged with verifying that the resulting systems did not discriminate along protected characteristics, giving engineers opportunities to retrain their systems before deployment if necessary. Where systems cannot be made to operate in a non-discriminatory fashion, the Recordkeeper would have the capacity and responsibility to detect that defect and prevent the system from being deployed.

This hypothetical Recordkeeper would be best realized as a new AI-focused government agency, which would maintain these databases, interface with the corporations that will use them, and evaluate the ML systems they develop.[234] This agency could ensure that the best possible datasets were being used and that engineers were taking every appropriate step available to them to counter the risk of discriminatory algorithms. Importantly, this solution would not necessarily forbid the deployment of algorithms that may be biased; it would only ensure that all appropriate steps were taken to mitigate the risk of their being biased in a way humans are not

---

[232] Though outside the scope of this Note, researchers have been developing techniques to train ML systems without exposing the data they are trained upon. *See, e.g.*, Martín Abadi et al., *Deep Learning with Differential Privacy*, 2016 PROC. ACM SIGSAC CONF. ON COMPUTER & COMM. SECURITY 308, 308–09.

[233] *See supra* note 84 and accompanying text.

[234] This Note is not the first piece of scholarship to propose an agency dedicated to regulating ML systems. *See generally* Andrew Tutt, *An FDA For Algorithms*, 69 ADMIN L. REV. 83 (2016) (arguing that the problem of machine bias is sufficiently pervasive in ML systems and can be expected to arise in sufficiently varied contexts that having an agency with expertise dedicated to promulgating standards for detecting and addressing such bias may be desirable). Tutt's argument holds especially true if detecting such bias requires a test set of private information against which to verify the system's performance: the less disseminated such a dataset would be, the better.

This Note goes further than Tutt in calling for such an agency to maintain the kind of database that is necessary to avoid creating such biased systems. Furthermore, while, like Tutt, this Note favors a government agency due to the government's accountability to its people, the Recordkeeper need not be a government agency; a nonprofit, public–private venture, or even industry consortium could fill this role.

permitted to be and would seek to understand how exactly these algorithms are biased.

Several criticisms of this proposal could be raised. The first concerns feasibility and enforceability: the creation of a new agency is too costly and presents an unrealistic goal, and any non-governmental solution will simply be ignored. But after start-up costs to gather the initial datasets and create the necessary technological infrastructure (including security infrastructure), the Recordkeeper could raise the funds necessary to manage itself and to keep its dataset current by charging fees to those who use it.[235] As for non-governmental Recordkeepers, companies might wish to signal to consumers that they worked to debias their systems, much as building developers tout LEED certification.[236] Regardless, this Note aims to describe a long-term solution to the problem of algorithmic discrimination; practical details of exactly how a Recordkeeper would be structured and funded fall outside its scope.

A second criticism is that to take this proposal to its extreme is tantamount to outlawing DLNNs, despite this Note's assurances to the contrary. But that would only be true if it were entirely impossible or unfeasible to mitigate the risks that such systems will discriminate. The work on bias mitigation and related areas indicates that such fears are misguided.[237] And the solution this Note calls for does not forbid the deployment of biased systems so long as every reasonable step was taken to avoid having that system be biased, meaning that such systems comport with, at minimum, the standards to which a human would be held.

Critics may note that ML system developers often have specialized data their systems must train upon, and the Recordkeeper's dataset will not include those data. But if the people whose data comprise the dataset provide that information voluntarily and for compensation, the developers can pay those same people to provide the specialized data the developers need to train their system, or conversely the developers can encourage the people whose data they already have to sign up with the Recordkeeper. If this transaction occurs with the Recordkeeper as an intermediary, the developers need not

---

[235] *See generally* U.S. DEP'T OF AGRICULTURE, AGRICULTURAL ECONOMIC REPORT NO. 775, USER-FEE FINANCING OF USDA MEAT AND POULTRY INSPECTION 6 (1999) ("Many Federal agencies now rely on user fees for at least some funding, and the importance of user fees as a source of funding has grown sharply in recent years.").

[236] *See, e.g.*, Daniel C. Matisoff et al., *Performance or Marketing Benefits? The Case of LEED Certification*, 48 ENV. SCI. & TECH. 2001, 2001 (2014) (discussing the importance of marketing-based benefits brought by Leadership in Energy and Environmental Design (LEED)-certified buildings because they indicate a building is "green").

[237] *See supra* Section I.C.2.

necessarily even know who those people are, and certainly need not access any of the private information about those people held by the Recordkeeper.

Finally and unfortunately, this solution would not necessarily address the problem of datasets encoding the systematic human and societal biases that shape their data. This is a separate problem entirely.[238] But at least a large, representative dataset is unlikely to encode random bias artifacts that may appear in smaller datasets and will be less vulnerable to malicious actors manipulating the dataset to encode those actors' own biases. While concededly not a panacea, this Note's proposal thus represents a significant improvement over the status quo.

In the advertising domain in particular, there is a simple economic argument for requiring companies like Facebook to prove that their advertising systems are not discriminatory before deploying them. Unlike hiring decisions, advertising decisions are relatively low cost. Advertisers may pay a flat fee for the ad, plus a small amount per click.[239] Therefore, it is fairly cheap to display an employment ad to additional people who may, in the end, not be qualified to perform the job in question. This low cost of compliance suggests that the standard for justifying disparate treatment of protected groups should be as strict as possible and should be imposed on both the ad platform (e.g., Facebook) and the advertiser.[240] The advertiser will lose only the money from paying for clicks by unqualified people; Facebook will lose only money from showing ads to people who do not click them (thereby losing the revenue Facebook would gain from showing users ads they would click). These costs are extremely low relative to those involved in interviewing and possibly hiring the wrong candidates. Most importantly, this will preserve everyone's incentives to accurately predict who will be interested in and qualified for an advertised job opportunity while ensuring that ads are not placed in a discriminatory fashion. Such a model will incentivize employers to provide useful nondiscriminatory characteristics to Facebook, and will incentive Facebook only to expand the pool of targeted users to those who will click the ad.

In summary, because preventing algorithmic discrimination requires a proactive solution, this Note proposes the creation of a Recordkeeper, an institution to assemble, manage, and put to use a large, rich, representative dataset of real human data. The Recordkeeper would allow ML developers limited and controlled access to that data, would audit those systems to

---

238 *See, e.g.*, Angwin et al., *supra* note 61.

239 Latanya Sweeney, *Discrimination in Online Ad Delivery*, 11 ACM QUEUE 1, 7 (2013).

240 SHAVELL, *supra* note 202, at 180.

ensure appropriate steps were taken to avoid learning biased functions, and would verify that their performance is as unbiased as possible.

CONCLUSION

Machine learning techniques, especially deep learning neural networks, pose new challenges for employment discrimination and other forms of discrimination. Employment discrimination law as it currently operates is ill-equipped to deal with the challenges posed by such systems. Nonetheless, scholars have proposed adaptations to the legal landscape to allow plaintiffs to bring lawsuits when they have been discriminated against by biased algorithms. But because of how DLNNs work, fixing employment discrimination litigation is unlikely to prevent such discrimination from occurring in the first place. The best chance at preventing discrimination is to ensure that DLNNs are trained on a large, representative dataset and precleared by an organization with expertise before they are deployed onto unsuspecting users.

The solution called for herein is ambitious. Even the reactive changes that courts could take without the legislative action called for will likely meet with resistance. The Supreme Court has been reticent to overly expand access to claims based on disparate impact, arguing that "disparate-impact liability must be limited so employers and other regulated entities are able to make the practical business choices and profit-related decisions that sustain a vibrant and dynamic free-enterprise system."[241] And the regulated parties themselves can be expected to resist such changes. Employers would rather only target ads to the people they think are most likely to qualify for the job, even if such people are defined on the basis of a protected characteristic; advertising platforms don't want to open their systems up for inspection.

But big data and machine learning-based decision-making are here to stay. They pervade every part of our lives and society.[242] These systems are fundamentally new in that they defy inspection, explanation, and

[241] Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Proj. Inc., 135 S. Ct. 2507, 2518 (2015).

[242] A version of the problems this Note has described will be presented by nearly any automated ML system that operates in the real world. Self-driving cars must accurately analyze their surroundings, systems like IBM's Watson are already making and will increasingly make business and healthcare analyses and recommendations, automated messaging centers must understand and accurately respond to callers regardless of accent or dialect, image-labeling systems may be trained extensively on images of people of one race and thus perform poorly on images of people of another race, to say nothing of hiring systems, credit rating systems, surveillance systems, and more. This warning is not to be alarmist—it is wonderful that these systems will alleviate human workload burdens, never get upset, and never suffer attentional lapses. But these systems are only as good as their design and their training data. Ensuring their training data is truly representative of the world in which the systems will operate and that they are designed to anticipate and overcome the pitfalls that may face them will be one of the main projects of the twenty-first century.

comprehension, at least beyond examining the data upon which, and the manner in which, the algorithms are trained. Just as the rise of industrialization required a new legal doctrine that is now widely accepted and part of the fabric of our legal lives (that is, principles of strict product liability), so too the rise of big-data algorithmic decision-making will require the legal system to address inscrutable decisions based on extrapolating patterns and data. These decisions are made without intent and defy careful causal analysis: they are something new under the sun. If we do not adapt our legal systems to them, we may be able to react on a case-by-case basis to their discrimination but will fail to prevent such discrimination from occurring in the first place. The algorithms will search for their signals, and the law will be lost in the noise.