(Draft in Progress, comments welcomed)
To be presented at the Workshop: Black Box Artificial Intelligence and the Rule of Law
Duke School of Law, January 22, 2021

NEGLIGENT ALGORITHMIC DISCRIMINATION

ANDRÉS PÁEZ*

I
INTRODUCTION

During the past few years, most of the largest companies in the world have started using various types of supervised machine learning algorithms in the sourcing, screening, interviewing and selection of candidates for a job.[1] Despite being promoted as a reliable means to eliminate bias,[2] these algorithmic decision-making tools have raised legal and moral questions because of their unlawful discriminatory effects in every step of the hiring process. The training phase of the machine learning systems used in these processes has been identified as the main source of bias. Algorithms tend to reproduce and even exacerbate the biases present in the trainer's mind or encoded in the datasets used to train the system. Some of this training bias is intentional and some is not. The question of intention is central from a legal perspective. Intentional discrimination is usually analyzed using the doctrine of disparate treatment, while non-intentional but avoidable discrimination is treated using the doctrine of disparate impact.[3] Proving that an algorithm has been intentionally trained to be biased is difficult. It requires either proving that an employer unlawfully classified individuals according to their membership in a protected class under Title VII of the 1964

[*] Associate Professor of Philosophy and member of the Center for Research and Formation in Artificial Intelligence, Universidad de los Andes, Bogotá, Colombia.

[1] Jon Shields, *Over 98% of Fortune 500 Companies Use Applicant Tracking Systems (ATS),* JOBSCAN (June 20, 2018), https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems/.

[2] Tomas Chamorro-Prezumic & Reece Akhtar, *Should Companies Use AI to Assess Job Candidates?* HARVARD BUSINESS REVIEW (May 17, 2019), https://hbr.org/2019/05/should-companies-use-ai-to-assess-job-candidates.

[3] These terms were first introduced in *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971).

Civil Rights Act,[4] or circumstantial evidence that discrimination was the main cause of an unfavorable employment decision. In the former case, part of the problem stems from the fact that vendors of algorithmic screening tools rarely disclose details about the construction and validation of the methods used, typically because they are proprietary and contain private, sensitive data.[5] But even if vendors were required to provide full access to their datasets and algorithms, they have other means at their disposal to mask unlawful discrimination using lawful proxy labels to classify data.[6] In the latter case, proving discriminatory intent from stray remarks or other circumstantial evidence is challenging even in hiring processes that do not use AI technology.

Under the doctrine of disparate impact, a plaintiff must show that a given practice disproportionally excludes a group protected by the Civil Rights Act. The employer can retort that despite its discriminatory impact the hiring practice is necessary to the essential operation of a business. Even if this is established, the plaintiff may still be successful by showing that the employer could have used an "alternative employment practice" with less discriminatory results. It is at this stage that the plaintiff's chances of success become slim in the context of AI given the black box nature of the algorithms employed.

Given the probatory obstacles of establishing disparate treatment and disparate impact, this paper explores the possibility of approaching the problem of algorithmic discrimination as a case of negligence. Some authors have argued that employers have the duty to protect candidates and employees from discriminatory treatment, and that an employer's failure to exercise due care in the manner of choosing employees is a violation of that duty. The issue turns on whether an employer can exercise due care in training the decision models used in the hiring process. The plaintiff must show that the breach of care caused the discriminatory effect and that this effect was reasonably foreseeable. Both of these requirements are a tall-call in the context of black box models that are often opaque and are not easily understood

[4] 42 U.S.C. §§ 2000e to -17 (1988). Title VII prohibits employers of 15 or more persons from discrimination in employment on the basis of race, color, religion, sex, or national origin.

[5] Manish Raghavan, Solon Barocas, Jon Kleinberg & Karen Levy, *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices*, PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (2020).

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

in causal terms. Nonetheless, a reinterpretation of the foreseeability condition might make this interpretation more plausible.

I will begin by offering some recent examples of algorithmic discrimination in hiring decisions that have been documented in recent years. In Part III, I explain why the doctrines of disparate treatment and disparate impact are ineffective in the case of algorithmic discrimination. Part IV explains how discrimination generally can be understood as negligence and examines whether this approach can be transferred to the context of algorithmic discrimination. I conclude by suggesting another novel way of analyzing algorithmic discrimination.

II
RECENT CASES OF ALGORITHMIC DISCRIMINATION IN HIRING DECISIONS

Machine learning algorithms are used in the sourcing, screening, interviewing, and selection of candidates. In this section I will provide specific examples of how these algorithms are used in each of these four stages and I will present recent studies that document their discriminatory effect.

Job offers are promoted on the internet using advertisement platforms such as Google Ads and social media such as Facebook. Advertisers are forbidden by law to choose the target audience of their ads using categories such as race, religion, sex, or national origin. Nonetheless, these algorithms select the target audience of job offers using browsing histories and viewing preferences, resulting in an inequitable distribution of the offers. A series of 21 experiments performed by researchers at Carnegie Mellon using 17,370 artificial agents collected over 600,000 real ads. Several of the experiments revealed that Google Ads tends to show the best paying jobs to males in higher proportion than females.[7] Ali et al.[8] found

[7] Amit Datta, Michael Carl Tschantz & Anupam Datta, *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination*. PROCEEDINGS ON PRIVACY ENHANCING TECHNOLOGIES 2015 (2015).

[8] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove & Aaron Rieke, *Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes*. PROCEEDINGS OF THE ACM ON HUMAN-COMPUTER INTERACTION (2019).

similar results in an experiment consisting of three ads placed on Facebook following their non-discriminatory policy: "Ads must not discriminate or encourage discrimination against people based on personal attributes such as race, ethnicity, color, national origin, religion, age, sex, sexual orientation, gender identity, family status, disability, medical or genetic condition."[9] Nonetheless, the three ads were shown to users based on their race and gender.

Companies also use resume search engines, which are tools that allow recruiters to search for candidates based on keywords and filters. Chen et al.[10] examined the algorithms used by the firms Indeed, Monster and CareerBuilder.[11] They ran queries on each site's resume search engine for 35 job titles across 20 American cities. Their final dataset included over 855,000 job candidates. They found that these resume search engines produce rankings that exhibit individual-level and group-level gender-based inequalities. The authors define the "individual fairness" of an algorithm as its capacity to place candidates with similar features at similar ranks; and "group fairness" as the assignment of similar distributions of ranks to men and women. The search engines exhibited significant group unfairness, but the size of the gender effect was small in terms of individual fairness. Similar biases have been found in TaskRabbit y Fiverr,[12] two resume search engines that focus on freelance workers.[13]

It has become common for candidates to be initially interviewed by chatbots. *Mya*, *Olivia*, *Myra* and *Yva* are among the best-known bots.[14] More sophisticated programs, such as the ones used by the firm HireVue,[15] analyze the speech patterns and facial expressions of job seekers as they answer personal and job-related questions on camera. According to the firm, each minute of video provides up to half a million data points that are analyzed by a

[9] Facebook, *Advertising Policies: 3. Discriminatory Practices* (2020), https://www.facebook.com/policies/ads/prohibited_content/discriminatory_practices#

[10] Le Chen, Ruijun Ma, Anikó Hannák & Christo Wilson, *Investigating the Impact of Gender on Rank in Resume Search Engines*, ANNUAL CONFERENCE OF THE ACM SPECIAL INTEREST GROUP ON COMPUTER HUMAN INTERACTION (2018).

[11] www.indeed.com, www.monsters.com, www.careerbuilder.com.

[12] www.taskrabbit.com, www.fiverr.com.

[13] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier & Christo Wilson, *Bias in Online Freelance Marketplaces: Evidence from Taskrabbit and Fiverr*, PROCEEDINGS OF THE 2017 ACM CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK AND SOCIAL COMPUTING (2017).

[14] hiremya.com, www.olivia.ai, www.myralabs.com, yva.ai.

[15] www.hirevue.com.

machine learning algorithm that detects traits such as emotional intelligence and social abilities. HireVue also requires cognitive and neurological tests, some of which have a videogame format. According to the firm, the results provide companies with information about the work style of the candidates, their willingness to learn, their ability to work with others, their personality and general cognitive abilities, their conscientiousness and responsibility, all of which are soft skills that cannot be easily deduced from a resume.

In 2019, the Electronic Privacy Information Center (EPIC), a rights group, filed an official complaint calling on the Federal Trade Commission to investigate HireVue's business practices. They argued that the system's "biased, unprovable and not replicable"[16] results constitute a major threat to American workers' privacy and livelihoods. One of the complaints was related to the use of videogames, which puts older candidates at a disadvantage. But the main complaint was directed at the algorithm that analyzed the videos recorded by the candidates. According to HireVue, 29% of a candidate's score was based on facial movements.[17] EPIC found this inadmissible: "The eye movement tracking captured in video assessments could discriminate against candidates with neurological differences. Eye movement tracking technology can be used to diagnose autism, Parkinson's, Alzheimer's, and psychiatric conditions like depression. Individuals with Autism Spectrum Disorder tend to look at people's mouths rather than making eye contact."[18] Furthermore, HireVue lacks a "reasonable basis" to support this technology and is therefore engaged in a deceptive trade practice in violation of the Federal Trade Commission Act.[19] In fact, many scientists are highly skeptical that an algorithm can correctly infer emotions and an individual's personality from facial expressions.[20]

[16] Electronic Privacy Information Center, *In Re HireVue* (November 6, 2019), at 7.

[17] Drew Harwell, *A Face-Scanning Algorithm Increasingly Decides whether You Deserve the Job*, THE WASHINGTON POST (October 22, 2019), https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/.

[18] Electronic Privacy Information Center, *supra* note 16, at 7.

[19] *Id.* at 9.

[20] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez & Seth D. Pollak, *Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements*, 20 PSYCHOLOGICAL SCIENCE IN THE PUBLIC INTEREST 1 (2019).

Finally, the most straightforward use of AI in labor selection is the use of resume-reading algorithms. Famously, Amazon abandoned an AI recruiting tool it developed because of its gender bias. Its purpose was to give job candidates scores ranging from one to five stars, much like consumers rate products on Amazon. The system was trained using the resumes of the company's employees over a 10-year period. As is well known, the tech industry has a notable gender imbalance that was reflected in the data set. In consequence, the system initially penalized resumes that included the word "women's" and the names of two all-women's colleges. After this glitch was corrected, the system found other ways to infer the applicant's gender and the whole project was scratched.[21]

This is just a small sample of discriminatory practices in labor decisions.[22] It is a hodgepodge of different types of decision systems. Despite their heterogeneity, they share at least five features in common: (i) the training data set and the structure of the model are often industrial secrets or protected by privacy laws; (ii) most of these systems are black boxes whose decisions are not readily explainable; (iii) it is easy to mask discrimination against a protected class using proxy properties; (iv) in the US there are plenty of legal obstacles for conducting research on algorithmic discrimination in many of these platforms; and (v) conflicts between individual and group fairness may arise when the algorithms are deployed. I will address these obstacles in more detail in the following sections.

III
Disparate Treatment and Disparate Impact

When Congress enacted Title VII of the 1964 Civil Rights Act it did not establish the standard that courts should require for proof of discrimination. Two Supreme Court decisions filled the void. In one case, the Court permitted a strict liability test similar to that used in

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

[22] Barocas & Selbst, *supra* note 6, and Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 Wm. & Mary L. Rev. 857 (2016) present many other examples of algorithmic discrimination in labor decisions.

strict liability in tort;[23] in the other, it required an intent test equivalent to the standard of proof for intentional torts.[24] These decisions gave rise to the doctrines of disparate impact and disparate treatment, respectively. The subsequent complete "tortification" of employment discrimination law has been amply documented.[25]

A strict liability test implies that a plaintiff must show that a given practice disproportionally excludes a group protected by the Civil Rights Act. To make a prima facie case of disparate impact, the Uniform Guidelines on Employment Selection Procedures[26] provide the so-called "four-fifths rule." It states that the selection rate for a protected group cannot be less than four-fifths that of the group with the highest rate. Often, compiling the requisite statistics to show that the policy has a disparate impact is costly and difficult, which imposes on the plaintiff a technical and financial burden. If the plaintiff is successful in establishing disparate impact as an initial matter, *Griggs* provides an affirmative defense for the employer. If a practice or hiring method is necessary to the essential operation of a business, it can be used despite its discriminatory impact. The so-called "business necessity defense" imposes on the employer the burden of showing that any job requirement that has a differential impact must have a manifest relationship to the employment in question. Courts apply different standards of relevance to the job-relatedness of the job requirement, but in general "courts tend to accept most common business practices for which an employer has a plausible story."[27] Finally, if the business necessity defense is successful, the plaintiff can return with proof that there was an alternative employment practice that the employer refused to use, "but which was equally effective in the business objective and less discriminatory."[28] In brief, there is a significant burden of proof placed on the job candidate, and the data reflects this burden: Plaintiffs of disparate impact cases only had on average a 19.2% success rate in

[23] *See, Griggs*, 401 U.S. 424. Although the court did not use the term "strict liability," the test was equivalent in practical terms.

[24] *McDonnell Douglas Corp. v. Green*, 411 U.S. 792 (1973).

[25] *See*, Charles A. Sullivan, *Tortifying Employment Discrimination*, 92 B. U. L. REV. 1431 (2012); Sandra F. Sperino, *Let's Pretend Discrimination Is a Tort*, 75 OHIO ST. L. J. 1107 (2014).

[26] Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607.4(D)(2015).

[27] Barocas & Selbst, *supra* note 6, at 707.

[28] 42 U.S.C. § 2000e-2(k)(1)(A)(2018).

seven years between 1984-2001 for Court of Appeals decisions, and a 25.1% success rate in six years between 1983-2002 for District Court decisions.[29]

While disparate impact focuses on systemic group discrimination, disparate treatment is used to establish whether an individual was discriminated. Disparate treatment comprises two different strains of discrimination: formal disparate treatment of similarly situated people and intent to discriminate. The former corresponds to an employer who unlawfully classifies individuals according to membership in a protected class, regardless of its reasons to do so; the latter is more closely associated with a conscious prejudiced attitude towards a protected class. Circumstantial evidence can be used to prove that discrimination was the main cause of an employment decision. Disparaging remarks made by the employer or procedural irregularities in promotion or hiring count as clear evidence of ill intent, but absent these elements, finding intent from stray remarks or other circumstantial evidence is challenging. Disparate treatment cases can also be tried under the mixed-motive framework, first recognized in *Price Waterhouse v. Hopkins*.[30] A plaintiff need not demonstrate that he was intentionally discriminated, but only that discrimination was a "motivating factor." This latter phrase has been interpreted by some to allow unconscious prejudice to be included under the disparate treatment regime.[31] However, not everyone agrees that discrimination due to unconscious bias should be considered disparate treatment.[32]

Now, proving disparate treatment in the case of algorithmic discrimination is unfeasible. Many of the decisions involved in training an algorithm are the result of the trainer's implicit prejudices, which in turn may be a reflection of cultural stereotypes prevalent in his social environment. Most discrimination that arises in data mining is thus unintentional.[33] In the few cases in which the trainer intentionally uses a discriminatory

[29] Michael Selmi, *Was the Disparate Impact Theory a Mistake?* 53 UCLA L. REV. 701, 738-739 (2005).

[30] *Price Waterhouse v. Hopkins*, 490 U.S. 228 (1989).

[31] *See*, Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. LAW REVIEW 1161 (1995).

[32] *See*, Charles A. Sullivan, *Disparate Impact: Looking Past the* Desert Palace *Mirage*, 47 WM. & MARY L. REV. 911 (2005).

[33] The question of the moral and legal responsibilities for implict biases has been the focus of much recent research both in philosophy and law. *See*, *e.g.*, Angela M. Smith, *Responsibility for Attitudes: Activity and Passivity in Mental Life*, 115 ETHICS 236 (2005); Christine Jolls & Cass R. Sunstein, *The Law of Implicit*

predictive model, equal protection restrictions are easy to circumvent using proxy categories that represent race, color, religion, sex, or national origin. Proving that this masking was ill-intentioned is difficult, if not impossible, because employers can always disclaim any knowledge of the proxy manipulation. In addition, Barocas and Selbst argue, "it is also possible to intentionally bias the data collection process, purposefully mislabel examples, or deliberately use an insufficiently rich set of features ... These methods of intentional discrimination will look, for all intents and purposes, identical to the unintentional discrimination that can result from data mining."[34] In sum, they conclude that "disparate treatment doctrine does not appear to do much to regulate discriminatory data mining."[35]

Turning now to disparate impact in the context of algorithmic discrimination, we find that some probatory challenges stay the same, but others become more difficult. *Prima facie*, the four-fifths rule that proves disparate impact remains unaltered by the change of context, as well as the business necessity defense. To prove that a job requirement that has a differential impact has a manifest relationship to the employment in question, it is indifferent whether the requirement was stated in a job ad or programed as a target variable in an AI decision model. Disparate impact liability can be found if the target variables are improperly chosen.[36]

But now consider the probatory difficulties that arise from the plaintiff's burden to prove that there was an alternative, less discriminatory, and equally effective employment

Bias, 94 CALIF. L. REV. 969 (2006); Linda Hamilton Krieger & Susan T. Fiske, *Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate Treatment*, 94 CALIF. L. REV. 997 (2006); Jules Holroyd, Robin Scaife & Tom Stafford, *Responsibility for Implicit Bias*, 43 JOURNAL OF SOCIAL PHILOSOPHY 274 (2012); Neil Levy, *Implicit Bias and Moral Responsibility: Probing the Data,* 94 PHILOSOPHY AND PHENOMENOLOGICAL RESEARCH 3 (2017). The discussion of this issue would take us too far afield.

[34] Barocas & Selbst, *supra* note 6, at 712.

[35] *Id*. at 701.

[36] Some vendors make sure that their models comply with the 4/5 rule so their clients do not have to demonstrate business necessity. But complying with the rule does not guarantee that the model does not have discriminatory effects. Vendors also have to take into account differences in model accuracy across the population. If the quality of its predictions differs dramatically between groups—a phenomenon known as differential validity—the model sets a group of people up to fail, reinforcing negative stereotypes. *See*, Raghavan et al., *supra* note 5, at 13.

practice that the employer refused to use. An alternative in this context would be a better decision model. This would be an unsolvable problem in the case of third-party models, like the ones used by Facebook or HireVue, because the employer who hires these firms does not have the possibility of tinkering with their models. This obstacle is removed if the company uses its own model,[37] but even in that case it is not obvious that the plaintiff will be able to show that a better alternative exists. Barocas and Selbst are skeptical that this is feasible in algorithmic discrimination cases. Most of the algorithms considered in the previous section are black boxes, so it is not possible to know with any degree of certainty whether a different or larger dataset, or the choice of different labels, would produce a less discriminatory outcome. The probatory obstacles are the same as in the case of masking. In the end, they conclude, "disparate treatment and disparate impact become essentially the same thing from an evidentiary perspective."[38]

Given the plaintiff's apparently insurmountable barriers in proving that there was an alternative employment practice under the disparate impact doctrine, a new alternative has to be found to provide relief to discriminated members of protected classes. In the remaining pages of this paper, I explore whether it is viable for a plaintiff to prove that a company breached its duty not to harm others, which is an essential element of a negligence claim. The doctrine of negligence might be better suited than disparate impact to address the recent advances in algorithmic hiring tools.

IV
ALGORITHMIC DISCRIMINATION AS NEGLIGENCE

As we saw in the previous section, the Supreme Court has often invoked tort common law to interpret federal discrimination statutes, a trend that has intensified in recent years. Although the trend to "tortify" federal anti-discrimination law is interpreted by some as an

[37] To simplify the analysis of employer liability, this is the only option I will consider from here onwards. Whether there can be and should be Title VII liability for vendors is left as an open question.

[38] Barocas & Selbst, *supra* note 6, at 713. Kim, *supra* note 27, at 910 argues that these two doctrines do not exhaust the options for demonstrating the discrimination forbidden by Title VII, as Barocas and Selbst seem to assume.

attempt to restrict it "by tightening causal standards,"[39] it also opens possible alternatives to explore algorithmic discrimination, especially since the doctrines of disparate treatment and disparate impact seem to be ineffective in this case. In this section I will explain why discrimination can be understood as negligence, and I will examine whether this analysis can be transferred to the context of algorithmic discrimination.

Oppenheimer[40] argues that Title VII discrimination can be interpreted according to a theory analogous to the third major doctrine of tort law, the doctrine of negligence.[41] Negligence is a breach of our duty to protect others, and in his view an employer's failure to exercise due care in the manner of choosing employees, or maintaining or terminating their employment, is a breach of that duty. In particular, an employer who does not make sure to use the least discriminatory employment practice available is acting negligently:

> Liability is established because the employer could have provided greater protection against discrimination without sacrificing its legitimate and necessary business interests. If a less discriminatory alternative exists, the employer has failed to act reasonably—it has breached its duty of care—by engaging in avoidable discrimination.[42]

Liability thus turns on what the employer knows or should have known about the risk of harm—the discriminatory effects—of its practice and on in its ability to prevent it. The employer's liability is not the result of an intent to discriminate, as in disparate treatment, nor is the employer strictly liable, as in disparate impact. Negligence thus offers a third possible analysis of discrimination.

Hart and Honoré's characterize negligence in the following terms: "A defendant is responsible for and only for such harm as he could reasonably have foreseen and prevented."[43] The plaintiff must show that the breach of care *caused* the discriminatory

[39] Sperino, *supra* note 25, at 1107.

[40] David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899 (1993).

[41] Oppenheimer's ideas have been further developed in Noah D. Zatz, *Managing the Macaw: Third-Party Harassers, Accommodation, and the Disaggregation of Discriminatory Intent*, 109 COLUM. L. REV. 1357 (2009); and Richard Thompson Ford, *Bias in the Air: Rethinking Employment Discrimination Law*, 66 STAN. L. REV. 1381 (2014). The title of the present paper is a nod to Oppenheimer's influential paper.

[42] Oppenheimer, *supra* note 40, at 933.

[43] H. L. A. HART & TONY HONORÉ, CAUSATION IN THE LAW (1985), at 255.

effect—and was thus preventable—and that this effect was reasonably *foreseeable*. The question is whether these two elements are applicable to the context of algorithmic discrimination.

Prime facie, the opaqueness of machine learning algorithms makes the causal and foreseeability clauses of negligence inapplicable. The plaintiff must prove that the defendant is aware of the particular causal mechanism that produces the discriminatory effect. Otherwise, the harm was neither preventable nor foreseeable. However, according to Selbst, "without interpretable or explainable AI, it is essentially impossible to claim that an AI error should have been foreseen ahead of time."[44] Another option for the plaintiff would be to prove that a richer or different set of input features would have generated a different output. Recent advances in post-hoc interpretability[45] would lend some plausibility to that strategy, which does not require any information about the causal structure of the model. However, it is not possible to know *a priori* what changes in the input features would generate a better outcome and it is unlikely that a court will force a defendant to undertake a costly and time-consuming revision of the model to compare different possible outcomes. Any procedure in that direction would open the gates to a wave of litigation that would overwhelm the industry. It seems, therefore, that there is no demonstrable level of care that a person can adhere to that would have prevented the harm.

Two alternatives present themselves at this juncture: either we abandon the idea of interpreting discrimination as negligence in the context of AI, or we reinterpret the notion of foreseeability in this new context and push for new legislation. I will argue that we should adopt the second prong of the dilemma and reinterpret what is foreseeable and reasonable within the limitations of black box algorithms. An effective legal response will require developing the doctrine of negligence to meet the particular challenges posed by data-driven discrimination.

[44] Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B. U. L. REV. 1315, 1362 (2020).

[45] *See*, *e.g.*, Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *"Why Should I Trust You?": Explaining the Predictions of any Classifier*, PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (2016).

The first point to consider is that we are discussing algorithmic discrimination in 2021, not in 2012. By now, there is plenty of evidence, some of it presented in Part II, that the models used in the different stages of hiring decisions are very likely to be biased. In a sense, discrimination has become foreseeable by default. The initial enthusiasm for hiring algorithms must give way to a modicum of prudence and caution. The fact that the probability that a new hiring model will be biased is high should put in motion anti-biasing protocols that become part of the duty of care of an employer. Instead of expecting the plaintiff in a disparate impact case to present evidence that there was an alternative, less discriminatory, and equally effective employment practice that the employer refused to use, the burden of proof ought to shift to the employer. Since most employment practices are discriminatory, it is the employer who must present evidence of its efforts to avoid bias. To prove a breach of the duty of care, the plaintiff need only show the absence of any precautionary measures. The harm only becomes unforeseeable when such measures have been implemented.

The second fact to consider is that there is a lack of transparency in the industry, which includes the non-disclosure of the models that have been discarded because of their discriminatory effects. The Amazon fiasco was discovered only because someone inside the company revealed the story to Reuters, but this is the exception that confirms the rule. An analogy with the pharmaceutical industry will be useful. Phase 1 trials provide the foundation for assessing the potential harm to humans of new experimental molecules. Such evidence is relevant not only to the harm profile of the particular molecule under investigation, but also to the harm profile of the class of molecules to which the particular molecule belongs.[46] Unfortunately, the vast majority of unsuccessful Phase 1 studies are not published, a phenomenon known as publication bias:[47]

> This publication bias of Phase 1 trials is wasteful. Future scientists who are unaware of the harm profile of *x* or other molecules of class *T*, for which prior Phase 1 trials have been performed, and who want to know the harm profile of *x* or another member of *T*, are liable to perform wasteful subsequent Phase 1 trials.[48]

[46] JACOB STEGENGA, MEDICAL NIHILISM 138 (2018).

[47] Evelyne Decullier, An-Wen Chan & François Chapuis, *Inadequate Dissemination of Phase I Trials: A Retrospective Cohort Study*, 6 PLOS MED e1000034 (2009).

[48] STEGENGA, *supra* note 46, at 138.

The most important consequence of publication bias, according to Stegenga, is that if one is unaware of the harm profile of a type of molecule, one's initial probability of its harm will be lower than it should, and so will its posterior probability, to speak in Bayesian terms. If molecules that appear safer than they should move on to phases 2 and 3, which do not focus on harm but on beneficial effects, the overall risk of harm in the general population increases.[49]

Transparency in the pharmaceutical industry would not only prevent much harm; it would also make it easier to determine civil liability when a company has been negligent in Phase 1 trials. A pharmaceutical company has a duty of care towards participants in an experimental trial, and failure to take into account similar Phase 1 studies with molecules of the same class would be a breach of care.[50] In a similar vein, transparency in the software industry is likely to have a beneficial effect by allowing engineers to learn from the mistakes of others. However, a recent study of how hiring algorithms are built, validated, and examined for bias reveals that most models and datasets are inaccessible to the public. Industry practices have to be gleaned or inferred from what companies publicly disclose.[51] Several authors have insisted on the need for independent audits of training data and model structure.[52] Perhaps the most elaborate proposal in that direction is the one by Langenkamp et al.,[53] of which I can only present an outline here. The authors introduce the idea of

[49] *Id.* at 139.

[50] Admitedly, establishing causation may be problematic, particularly for a research subject who is also a patient. In general, litigation for negligence in the investigator-subject relationship has not been very successful. *See*, Larry D. Scott, *Research Related Injury: Problems and Solutions*, 31 THE JOURNAL OF LAW, MEDICINE & ETHICS 419 (2003).

[51] *See*, Raghavan et al., *supra* note 5.

[52] *See*, *e.g.*, Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. 189 (2017); Julie E. Cohen, *The Regulatory State in the Information Age*, 17 THEORETICAL INQUIRIES IN LAW 369 (2016); Ifeoma Ajunwa, *The Auditing Imperative for Automated Hiring*, 34 Harv. J. L. & Tech. __ (forthcoming).

[53] Max Langenkamp, Allan Costa & Chris Cheung, *Hiring Fairly in the Age of Algorithms*, ARXIV PREPRINT arXiv:2004.07132 (2020). Other important work in that direction is Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III & Kate Crawford, *Datasheets for Datasets*, ARXIV PREPRINT arXiv:1803.09010 (2018); and Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa

"algorithmic transparency reports" that cover four categories: Intent (addresses the purpose of the model), Dataset (information about demographics, labels, and test sets), Metrics (measures of model performance, thresholds and definitions of "fairness"), and Applications (uses of the model in decision-making).[54] I would add that such reports should include a History category under which the employer must present the limitations and potential risks of previous versions of the model, and the measures taken to correct them.

These reports have a double function. On the one hand, they would be the factual basis for any claim about the existence or absence of precautionary measures that fulfill or breach the employer's duty of care, respectively.[55] On the other, access to the algorithmic transparency reports of a family of machine learning models would help prevent discriminatory effects. They have the same function as the reports of the failed Phase 1 trials of a new drug. Not taking them into account is a breach of reasonable care in both cases.

The causal element of negligence presents a bigger challenge. Tort law usually requires that a defendant's conduct was both the actual cause—but for which the harm would not have occurred—and the proximate cause—a reasonably foreseeable and not insignificant cause—of the harm. In the case of algorithmic negligence, the actual cause is an algorithm that has not been adequately tested for harmful effects. But for the omission to validate the non-discriminatory effects of the algorithm, the harm would probably not have occurred. Regarding the proximate cause, there is plenty of evidence that an algorithm that has not been subject to precautionary measures is very likely to produce a discriminatory effect. Thus, as I argued above, proving that the algorithm is *not* the proximate cause is part of the shift in the burden of proof from the plaintiff to the employer. The latter must show that it has taken every possible measure to eliminate the harmful, foreseeable effects of the model.

Deborah Raji &Timnit Gebru, *Model Cards for Model Reporting*, PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (2019).

[54] Incidentally, these reports could be regarded as "explanations" of the model.

[55] As noted by Raghavan et al., *supra* note 5, at 15, it might be impossible to apply any de-biasing methodology without using sensitive information about the protected classes to which people in the training data belong, but doing so can put employers in legal jeopardy for disparate treatment. A pressing challenge for algorithmic hiring is to find ways to solve the trade-off between protecting sensitive and private information and developing less discriminatory algorithms.

There have been similar proposals in the literature. Ifeoma Ajunwa has recently offered a theory of liability, which she calls the doctrine of "discrimination per se," to hold corporations accountable for algorithmic bias under Title VII. The basic idea is that "an employer's failure to audit and correct its automated hiring platforms for disparate impact should serve as prima facie evidence of discriminatory intent."[56] The theory reverses the American legal tradition of deference to employers by shifting the burden of proof:

> A plaintiff can assert that a hiring practice (for example, the use of proxy variables resulting in or *with the potential to result in* adverse impact to protected categories) is so egregious as to amount to *discrimination per se*, and this would shift the burden of proof from the plaintiff to the defendant (employer) to show that its practice is non-discriminatory.[57]

Inspired by a paper by Stephanie Bornstein on recklessness.[58] Ajunwa suggests that *negligence per se* should be the model for creating this new legal framework.[59] The introduction of a statutory auditing imperative imposed on the employer would be the basis for negligence per se liability.[60] However, Ajunwa's proposal runs against the grain of the current trend to move away from strict liability to a standard of reasonable care,[61] like the one presented in this paper. The path forward in my opinion is the adoption of non-mandatory industry standards established by an independent certifying entity. The standards adopted cannot be something akin to a checklist of formal requirements. New machine learning models are more akin to new molecular compounds whose harmful effects have to be detected experimentally. Without the evidence provided by something akin to a Phase 1 trial for machine learning models, and by algorithmic transparency reports, there is little that a certifying board would be able to assert with confidence about the model.

[56] Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671, 1672 (2019).
[57] *Id*. at 1728.
[58] Stephanie Bornstein, *Reckless Discrimination*, 105 CAL. L. REV. 1055 (2017).
[59] Ajunwa, *supra* note 56, at 1730.
[60] Ajunwa, *supra* note 52.
[61] Peter M. Gerhart, *The Death of Strict Liability*, 56 BUFF. L. REV. 245 (2008).

V
CONCLUSION

The use of algorithms in hiring decisions offers a well-defined setting to discuss algorithmic discrimination, but the foregoing analysis can be easily extended to other areas where algorithmic discrimination has been detected, such as racist ad targeting on Google Search[62] or sexist word associations in language models.[63] If any of these cases is to be treated as a case of negligence, it will require a legislative change in the law of torts that has not even been contemplated. This paper hopes to open that discussion at least at a theoretical level.

The paper also joins the call for transparency in machine learning and adds an additional call for experimental evidence about the safety of the models that affect people's lives, in the same way that we demand proof of the safety of drugs that have the potential to harm human subjects. Vendors of snake oil also fought to keep their proprietary formulas wrapped in secrecy until the public's interest prevailed. Vendors of hiring algorithms should be held to the same standard.

An even more radical possibility will be analyzed in future work. One way to hold a company accountable for algorithmic discrimination is to attribute some sort of personhood to the model and regard it as a negligently trained employee exhibiting implicit bias.[64] Negligent training claims arise when the employer incorrectly trains an employee and the employee's actions harm another individual. So far, harms generated by AI systems such as self-driving cars are usually analyzed in terms of product liability, instead of negligence, because there is no person who was negligently trained. But if we are willing to stretch our

[62] Latanya Sweeney, *Discrimination in Online Ad Delivery*, 11 QUEUE 10 (2013).

[63] Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 SCIENCE 183 (2017).

[64] *See*, Barocas and Selbst, *supra* note 6, at 699 ("Another option is to imagine the *model* as the decision maker exhibiting implicit bias. That is, because of biases hidden to the predictive model such as nonrepresentative data or mislabeled examples, the model reaches a discriminatory result."); Karni Chagal-Feferkorn, *The Reasonable Algorithm*, 1 J. L. TECH. & POL'Y 111 (2018) (identifies and addresses the conceptual difficulties stemming from applying a "reasonableness" standard on non-humans and the question of whether there is any practical meaning in analyzing the reasonableness of an algorithm separately from the reasonableness of its programmer).

concepts, liability for negligently trained or negligently supervised models becomes a possibility.