How should liability be attributed for harms caused by biases in Artificial Intelligence?

Alina Glaubitz

Senior Thesis, Yale Department of Political Science

Nathaniel Raymond

Advisor, Yale Jackson Institute for Global Affairs

April 29, 2021

Abstract

Artificial Intelligence (AI) tools are discriminating against minorities as a result of biases in their training data or programming. These tools include Amazon's hiring tool that discriminated against women, Facebook's algorithm that showed real estate ads to White users only, and the US government's facial recognition tools that misidentified Black women disproportionately more often than White men. Courts have struggled to attribute liability for discriminatory harms caused by AI. As AI-driven automated systems become the backbone of many industries, litigation will continue to arise about harms allegedly caused by those systems. Determining liability for algorithmic biases will become a critical step in safeguarding citizens from harm.

This paper argues for a theory of algorithmic liability within tort law rooted in Judge Hand's formulation of the duty of care. An algorithmic duty of care is proposed that incorporates a disparate impact assessment, determining the extent of any disproportionate, adverse impact on a protected class, and an assessment of alternative algorithmic solutions that could achieve the same legitimate purpose with a lesser impact on a protected class. This theory will be situated in existing case law and regulation, and the limitations of this theory will be discussed in relation to the current legal landscape of auditing and data protection.

Table of Contents

I. Introduction

Impact Pro, a healthcare algorithm that helps identify individuals who benefit most from population health management programs,[1] is applied to approximately 200 million patients across the United States every year, and predicts that Black patients have lower medical costs, suggesting that their illnesses are less severe.[2] The algorithm evaluates patients' medical history and medical spending to predict patients' future healthcare costs, ranking Black patients lower than white patients on their medical needs as a result of their lower medical spending.[3]

The alleged harms caused by the Impact Pro algorithm may soon become a test case for algorithmic liability in US tort law, the civil law of damages. Leticia James, Attorney General of New York, is reportedly investigating the impact of the algorithm's bias on Black patients.[4] The US legal system has yet to have a foundational precedent set for algorithmic liability. Regulatory systems are shaped by, among other inputs, tort litigation, and legislation is often enacted in response to court decisions, or in anticipation of court precedent. As AI-driven automated systems become the backbone many industries, litigation will continue to arise from harms allegedly caused by these systems. To date, a theory of algorithmic liability – an duty of care for algorithms – has not been developed in US jurisprudence.

How should liability be attributed for harms caused by biases in Artificial Intelligence? This question drives the following research. An algorithmic duty of care is argued for, rooted in a disparate impact assessment, determining the extent of any disproportionate, adverse impact

[1] OPTUM, "Impact Pro: Individual & Population Health Risk Analytics," OPTUM, last modified 2021, https://www.optum.com/business/solutions/data-analytics/data-analytics-health-plans/impact-pro-cpl.html.
[2] Allana Akhtar, "New York is Investigating UnitedHealth's Use of a Medical Algorithm That Steered Black Patients Away
[2] Allana Akhtar, "New York is Investigating UnitedHealth's Use of a Medical Algorithm That Steered Black Patients Away from Getting Higher-quality Care," Business Insider, last modified October 28, 2019, https://www.businessinsider.com/an-algorithm-treatment-to-white-patients-over-sicker-black-ones-2019-10.
[3] Melanie Evans and Anna W. Mathews, "New York Regulator Probes UnitedHealth Algorithm for Racial Bias," Wall Street Journal, last modified October 26, 2019, https://www.wsj.com/articles/new-york-regulator-probes-unitedhealth-algorithm-for-racial-bias-11572087601.
[4] Ibid.

on a protected class, and an assessment of alternative algorithmic solutions that could achieve the same legitimate purpose with a lesser impact on a protected class. Herein, race, color, religion, national origin, citizenship, sex (including gender, pregnancy, sexual orientation, and gender identity), age, physical or mental disability, familial status, veteran status, and genetic information, are defined as protected classes according to federal anti-discrimination laws.[5]

Black Americans spend less on healthcare on average due to a lack of access to medical services and a long-standing mistrust of the healthcare system.[6] This mistrust results, in part, from unethical research like the Tuskegee Syphilis study of the 1930s through 1970s, which deceived African-American test subjects and intentionally denied their access to treatment so that researchers could study the progression of syphilis.[7] A major consequence of the Impact Pro algorithm is the priorization of the care of healthier, white individuals over the care of sicker, Black individuals.[8]

Obermeyer et al. (2019) reverse engineered the Impact Pro algorithm and found that the algorithm's bias compromised the referral of Black patients for a higher degree of specialized healthcare.[9] Black patients were referred 17.7% of the time where they should have been referred 46.5% of the time.[10] "This compounds the already unacceptable racial biases that Black patients experience, and [health providers' and insurers'] reliance on such algorithms

[5] Civil Rights Act of 1964; Age Discrimination in Employment Act of 1967; Equal Pay Act of 1963; *Bostock v. Clayton County*; Pregnancy Discrimination Act; Civil Rights Act of 1968 Title VIII; Rehabilitation Act of 1973; Americans with Disabilities Act of 1990; Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act; Genetic Information Nondiscrimination Act.
[6] Akhtar, "New York is Investigating UnitedHealth's Use of a Medical Algorithm That Steered Black Patients Away from Getting Higher-quality Care."
[7] Centers for Disease Control and Prevention, "Tuskegee Study," CDC, last modified July 16, 2020, https://www.cdc.gov/tuskegee/timeline.htm.
[8] Evans and Mathews, "New York Regulator Probes UnitedHealth Algorithm for Racial Bias."
[9] Ziad Obermeyer et al., "Dissecting racial bias in an algorithm used to manage the health of populations," *Science* 366, no. 6464 (2019): 447, doi:10.1126/science.aax2342.
[10] Ibid.

appears to effectively codify racial discrimination as … policy," the New York Department of Health wrote in a letter to United Health, the owner of the Impact Pro algorithm.[11]

Algorithmic biases extend beyond healthcare into other sectors of society where nondiscrimination is protected by law, including in housing, education, employment, and criminal justice. Such biases are exemplified by Amazon's hiring tool that discriminated against women, Facebook's algorithm that showed real estate ads to only white users, and the US government's facial recognition tools that misidentified Black women disproportionately more often than white men.[12] AI tools have discriminated against minorities as a result of biases in their training data (data that algorithms learn from and make predictions on) or their programming, and courts have struggled to attribute liability for discriminatory harms caused by AI.

This paper draws largely from literature on algorithmic biases in political science, social science, law, and policy work. West, Whittaker and Crawford (2019) note that the use of AI systems for the classification, detection, and prediction of protected characteristics is problematic due to historically engrained biases.[13] Systems that use physical appearance as a proxy for character are harmful, including AI tools that claim to detect sexual orientation from headshots, predict criminality based on facial features, or assess employee competence via micro-expressions.[14] The authors note that such systems are replicating patterns of racial and gender bias in ways that can extend historical inequality, resulting in discriminatory harms.[15] Richardson, Schultz, and Crawford (2019) further demonstrate that historical biases feed into algorithmic biases, with law enforcement agencies increasingly using predictive policing

[11] Department of Financial Services, *Letter to CEO of UnitedHealth*, (New York: Department of Financial Services, 2019), https://dfs.ny.gov/system/files/documents/2019/10/20191025160637.pdf.
[12] Akhtar, "New York is Investigating UnitedHealth's Use of a Medical Algorithm That Steered Black Patients Away from Getting Higher-quality Care."
[13] West, S.M., Whittaker, M. and Crawford, K. Discriminating Systems: Gender, Race and Power in AI. *AI Now Institute* (2019): 3, https://ainowinstitute.org/discriminatingsystems.html.
[14] Ibid.
[15] Ibid.

systems to forecast criminal activity.[16] In numerous jurisdictions these systems are built on data that is racially biased or based on "dirty policing."[17] These historic practices and policies shape the environment in which algorithms are created, which raises the risk of producing inaccurate, skewed, or systemically biased outcomes.[18] This paper contributes to this growing body of literature in that it delineates how the law should attribute liability for the discriminatory outcomes prompted by biases in AI.

This attribution is based on an analysis of, and response to literature on algorithmic accountability, court precedent, state and federal regulations, and standards of liability within the technological field. A theory of algorithmic liability for harms caused by biases in AI will then be proposed within the framework of tort law and the duty of care.

II. Existing approaches to algorithmic accountability

Multiple approaches to algorithmic accountability exist, including the disclosure of source code and disparate impact assessments. In order to analyze these approaches, different typologies of AI need to be identified and discussed in relation to legal personhood and the conception of AI as a monolith.

II.i. Typologies of AI

The first generation of AI has been developed and the second generation of AI is now being developed. The first generation of AI is reactive – it reacts to input data based on a

[16] Richardson, R., Schultz, J. and Crawford, K. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *New York University Law Review Online* (2019): 15, https://www.nyulawreview.org/online-features/dirty-data-bad-predictions-how-civil-rights-violations-impact-police-data-predictive-policing-systems-and-justice/.
[17] Ibid.
[18] Ibid.

program or model.[19] The second generation of AI will have limited memory, reacting to input data by applying past experience to a novel environment.[20] Prototypes of the second generation of AI actively recall past scenarios or sets of stimuli related to the determination the algorithm is asked to make. The algorithm then weighs its options and based on its memory makes a decision in an unfamiliar environment. The third generation of AI will react to input data by making its own subjective evaluations under the theory of mind.[21] For example, it will speculate about *why* another machine or human is acting a certain way. Finally, the fourth generation of AI will be self-aware – aware of its 'machine-ness' and of its unprogrammed individual agency.[22] When an algorithm is able to *generate* the programming required to reflect by writing its own code, rather than be *given* the programming required to reflect, it has reached the fourth, self-aware stage of AI. This paper theorizes a duty of care for the first generation of AI – reactive artificial intelligence. As the development and deployment of subsequent generations takes place, the proposed duty of care will need to evolve in line with the new capabilities of algorithms.

AI will likely reach legal personhood between its third and fourth generations. That is, AI would be treated as a person for limited legal purposes and be able to enter into contracts, own property, or be sued. Lima (2018) argues that historically, the conception of personhood has been connected to the human ability to self-reflect and have a conscience.[23] In the context of criminal law, personhood is closely associated with responsibility, as only a person who can differentiate between right and wrong and is able to make choices can be held responsible for

[19] Michael P. Georgeff and Amy L. Lanksy, "Reactive Reasoning and Planning," Robotics, 1987, https://www.aaai.org/Papers/AAAI/1987/AAAI87-121.pdf.
[20] Bruce G. Buchanan, "A (Very) Brief History of Artificial Intelligence," *AI Magazine* 26, no. 4 (2005), https://ojs.aaai.org/index.php/aimagazine/article/view/1848.
[21] F. Cuzzolin et al., "Knowing me, knowing you: theory of mind in AI," *Psychological Medicine* 50, no. 7 (2020), doi:10.1017/s0033291720000835.
[22] Michael T. Cox, "Perpetual Self-Aware Cognitive Agents," *AI Magazine* 28, no. 1 (2007), https://doi.org/10.1609/aimag.v28i1.2027.
[23] Dafni Lima, "Could AI Agents Be Held Criminally Liable? Artificial Intelligence and the Challenges for Criminal Law," *South Carolina Law Review* 69 (2018): 684, https://www.researchgate.net/publication/335107356_Could_AI_Agents_Be_Held_Criminally_Liable_Artificial_Intelligence_and_the_Challenges_for_Criminal_Law.

choosing to do wrong.[24] The Model Penal Code defines criminal liability as "an offense…based on conduct which includes a voluntary act or the omission to perform an act of which he is physically capable."[25] An act is currently defined as a "bodily movement" (whether voluntary or not),[26] precluding algorithms from committing acts. Criminal liability depends on conceptions of AI as complex enough to perceive a situation and proceed with acting, or fail to act where it could have acted.[27] AI cannot yet be considered an agent that can understand its own significance as well as the relevance of its criminal conduct.[28] Similarly, people with diminished capacity (children and those suffering from mental illness, for example) are not subject to criminal sanctions due to their lack of self-awareness. Until AI has achieved self-awareness and individual agency, it likely cannot be considered a legal person or be held criminally responsible for its harmful conduct. As this paper theorizes liability for the first generation of AI, the proposed theory of algorithmic liability will remain within the framework of tort law as opposed to criminal law.

Algorithms can broadly be classified into the following categories of function: prioritization, classification, association, and filtering.[29] Harms stemming from bias can occur in each of these function categories. Algorithms prioritize information in a way that emphasizes certain elements at the expense of others; by definition, prioritization can result in discrimination by prioritizing the characteristis of certain classes over others.[30] Classification decisions mark a particular entity as belonging to a given class by considering key characteristics of that entity.[31] Class membership can then drive downstream discrimination.

---

[24] Ibid., 686.
[25] Ibid., 679.
[26] Ibid.
[27] Ibid., 684.
[28] Ibid., 689.
[29] Nicholas Diakopoulos, "Accountability in Algorithmic Decision Making," *Communications of the ACM* 59, no. 2 (2016): 57, doi:10.1145/2844110.
[30] Ibid.
[31] Ibid.

Association decisions require creating relationships between entities.[32] The similarity metrics that dictate how closely two entities match can affect the accuracy of an association and how that association is interpreted by others, for example an association between class and crime.[33] Filtering decisions involve including or excluding information according to rules or criteria.[34] Filtering has the potential to erase certain people, classes, or demographics.

Similarly, Mayson (2019) argues that current strategies that aim to eliminate bias in AI are "at best superficial and at worst counter-productive, because the source of racial inequality in risk assessment lies neither in the input data, nor in a particular algorithm, nor in algorithmic methodology per se."[35] The author contends that the problem stems from the nature of prediction itself, as all prediction draws on the past to estimate future events.[36] In a racially stratified world, any prediction method will project past inequalities into the future.[37]

Current approaches to the governance of AI are misguided by the conception of AI as a monolith. AI should be viewed as a series of automated steps toward an outcome, where these steps can cause harm individually as well as in aggregate. Bias can be encoded in an algorithm's component steps and in its totality. We cannot engage in the legal exercise of assigning liability without deconstructing AI to its component parts as well as analyzing its whole. For this reason, different types of audits are discussed in Section VII.ii. and VII.iv.

II.ii Reasonable person standard

Within tort law, many existing legal mechanisms for assigning tort liability depend on the reasonable person standard. Chagal-Feferkorn (2018) distinguishes the "reasonable person"

[32] Ibid., 58.
[33] Ibid.
[34] Ibid.
[35] Sandra G. Mayson, "Bias In, Bias Out," The Yale Law Journal 128 (2019): 2218, https://www.yalelawjournal.org/article/bias-in-bias-out.
[36] Ibid.
[37] Ibid.

standard from the "reasonable algorithm" standard.[38] The author questions whether one should analyze the reasonableness of an algorithm separate from the reasonableness of its programmer.[39] The author also highlights the potential legal implications of finding that an algorithm "acted" reasonably or unreasonably, and whether or not such an analysis reconciles with the rationales behind tort law.[40] The proposed duty of care for algorithms will not rely on a reasonable person standard in line with Chagal-Feferkorn's reasoning, because standards of reasonable conduct do not readily translate into the algorithmic realm with the primary perpetrator of harm being a non-human entity. As in the case of legal personhood, the reasonable person standard for algorithms may be introduced into tort liability once algorithms have evolved into the third and fourth generations of AI.

II.iii. Bias detection

Approaches to addressing algorithmic harms have predominantly relied on the detection of bias, often through deduction, retrospectively determining whether a dataset, model, output, or algorithmic system is biased. These include audits of training data, the disclosure of source code, as well as systems and operational audits (discussed in Section VII.ii).

Scholars have challenged the notion that bias can be identified and mitigated through the disclosure of an algorithm's training data and source code (Kroll et al., 2017; Ananny and Crawford, 2016). Kroll et al. (2017) argue that transparency of source code is not the solution to AI bias because compelling this disclosure may divulge proprietary information, may not be helpful for algorithms dependent on randomization or algorithms that change over time, and

[38] Karni Chagal-Feferkorn, "The Reasonable Algorithm," *Journal of Law, Technology and Policy* (forthcoming), written January 2018, 51, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3095436.
[39] Ibid.
[40] Ibid.

may allow individuals to "game the system."[41] To illustrate this final concern, the authors presented the example of the IRS examining tax returns for signs of tax evasion based on previously audited returns. If the public knows which variables on a tax return are considered signs of fraud, tax evaders may adjust their behavior and the original signs of fraud may lose their predictive value.[42] The authors conclude that AI regulation should seek procedural regularity (grounded in the Fourteenth Amendment's principle of procedural due process), in which each individual understands that the same procedure was applied to them and that the procedure was not designed to disadvantage them specifically.[43] Auditing through an independent data protection agency may be a means of establishing procedural regularity.

Diakopoulos (2016) recognizes, however, that the Freedom Of Information Act (FOIA) could compel the disclosure of government source code for public algorithms.[44] The author notes the example in which the Federal Highway Administration was required to reveal the source code of an algorithm it used to compute safety ratings for carriers, disclosing the weighting of factors used in that calculation.[45] While Kroll et al. argue against the disclosure of source code altogether, Diakopoulos observes additional hurdles to this solution to algorithmic bias. FOIA is limited to the algorithms of public agencies, and thus does not provide access to the proprietary information of a private sector algorithm. FOIA also does not require government agencies to create documents that do not already exist.[46] Hypothetically, a government algorithm could use a variable that corresponds to a protected class, such as race, and use that variable in determining its output. So long as that variable was never directly stored in a document, FOIA would not compel its disclosure.[47] The author suggests that audit

[41] Joshua A. Kroll et al., "Accountable Algorithms," *University of Pennsylvania Law Review* 165 (2017): 654, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2765268.
[42] Ibid., 654.
[43] Ibid., 692.
[44] Diakopoulos, "Accountability in Algorithmic Decision Making," 59.
[45] Ibid.
[46] Ibid.
[47] Ibid.

trails could help mitigate this issue by recording stepwise correlations and inferences made during an algorithm's prediction process, and guidelines should be developed for when government use of an algorithm would trigger an audit trail.[48] The author also proposes a Freedom Of Information Processing Act be established that would allow the public to submit datasets to the government for processing through its algorithm, and require the government to provide the algorithm's output.[49] That would allow interested parties, including journalists or policy experts, to run assessments that test government algorithms and look for cases of discrimination or censorship.[50] This proposal further highlights the importance of auditing.

In an attempt to assign tort liability to autonomous vehicles, Cowger Jr. (2018) argues that a victim of harm should be entitled to compensation without any finding of fault or responsibility.[51] The court system would be used to determine the amount of damages the victim would be entitled to, with the insurance sector paying out the compensation.[52] If the damages were disputed, courts could also turn to product liability law if the harm was not the result of an algorithmic decision, but rather a mechanical or software defect, or the intentional act of a third party.[53] Cowger Jr.'s conception of algorithmic liability in essence does not assign liability to any involved party. While it serves as an innovative solution to assigning liability for harm caused by autonomous vehicles, it does not transpose well into other sectors or AI applications that are uninsured.

[48] Ibid.
[49] Ibid.
[50] Ibid.
[51] Alfred R. Cowger, Jr., "Liability Considerations When Autonomous Vehicles Choose The Accident Victim," *Journal of High Technology Law* 14 (2018): 60, https://cpb-us-e1.wpmucdn.com/sites.suffolk.edu/dist/5/1153/files/2018/12/Cowger-FINAL-174f0gc.pdf.
[52] Ibid.
[53] Ibid.

II.iv. Disparate impact assessments

There is a growing body of literature arguing for disparate impact assessments of algorithms (MacCarthy 2017; Bornstein 2018). Bornstein (2018) highlights that the assumption has been made that algorithms are "facially neutral," where an algorithm does not appear to be discriminatory on its face; rather is may be discriminatory in its application.[54] In being facially neutral, algorithms pose no problem of unequal treatment. As a result, algorithmic discrimination cannot be challenged using an intent-based *disparate treatment* theory of liability under Title VII of the Civil Rights Act of 1964.[55] Instead, it presents a problem of unequal outcomes, subject to Title VII's *disparate impact* framework.[56] A disparate impact assessment of an algorithm would evaluate whether an algorithm is being used to substantially achieve a legitimate purpose, and the extent to which an algorithm disproportionately, adversely impacts a protected class.

Antidiscrimination law requires that people be treated equally, but it also requires that people be treated individually and not be judged against stereotypes associated with protected classes. Bornstein (2018) contends that when applying an anti-stereotyping lens to the problem of algorithmic discrimination, individuals are judged against a model that incorporates stereotypes of protected classes.[57] If an individual is algorithmically penalized for failing to conform to a stereotype, that may constitute intentional stereotyping. The author argues that having a computer execute subjective decision-making does not make otherwise biased action facially neutral.[58] Bornstein believes that framing algorithmic discrimination as actionable disparate treatment under a stereotyping theory may expand the reach of liability under

[54] Stephanie Bornstein, "Antidiscriminatory Algorithms," *Alabama Law Review* 70 (2018): 520, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3307893.
[55] Ibid.
[56] Ibid.
[57] Ibid., 571.
[58] Ibid.

current laws, enough to motivate the costly and complex efforts required to mitigate algorithmic biases.[59]

This paper's proposed duty of care will serve as a deductive tool that judges can apply to retroactively determine whether an algorithm is biased against a protected class. The growing body of literature on algorithmic accountability to which this paper hopes to contribute may also lead to the development of procedural regularity in this field.

III. Methodology

The research methodology of this paper relies on the database NexisUni in order to compile case law, state and federal regulations, as well as existing theories of liability.

With respect to case law, NexisUni was used to identify judgments containing the words "artificial intelligence" and filter these by narrowing down the search to cases that also included the terms "algorithm," "bias," and "minority." Cases involving gerrymandering were excluded, as there were numerous cases involving algorithmic simulations of voting districts that were not directly concerned with AI bias. A summary of cases compiled by the research institute *AI Now* in its "Litigating Algorithms 2019 US Report"[60] also informed the final selection of cases.

A similar process was followed to search for statutes concerning algorithms, filtering the NexisUni search results by "bill text," "artificial intelligence," and "bias." Regulations mitigating bias in artificial intelligence were geographically centered in the states of New Jersey, Washington, California, and Illanois. AI-based regulations compiled by Yoon Chae

[59] Ibid., 572.
[60] Rashida Richardson, Jason M. Schultz, and Vincent M. Southerland, "Litigating Algorithms 2019 US Report," *AI Now*, 2019, 28-31, https://ainowinstitute.org/litigatingalgorithms-2019-us.pdf.

(2020)[61] and the National Conference of State Legislatures[62] informed the final selection of regulations concerned with AI bias. While a multitude of regulations support the establishment of AI research and develop initiatives as well as oversight committees, only regulations directly concerned with AI bias were included in the review of existing regulation.

Law review articles and the work of scholars at the intersection of AI and law informed the analysis of existing theories of liability in the technology sector and specific to algorithms. The proposed algorithmic duty of care drew on disparate impact literature, as well as tort law's foundational precedent of *United States v. Carroll Towing Co.* and Judge Hand's formulaic conception of the duty of care.

IV. Regulation on AI bias

Legislation on AI bias over the past two years has worked towards prohibiting algorithmic discrimination by attempting to introduce impact assessments and establishing various research and policy initiatives to further the development of strategies to reduce AI bias. Of the following federal and state bills on AI bias, only one state bill (at the time of writing) was written into law. The United States has no federal regulation on AI bias. Existing regulation provides little guidance on how to identify and mitigate algorithmic biases, impeding accountability for the biased outcomes of AI.

IV.i. Algorithmic Accountability Act

The federal Algorithmic Accountability Act (AAA), introduced in Congress on April 10, 2019 through Senate and House bills S. 1108 and H.R. 2231. If the bill were to be enacted, it

[61] Yoon Chae, "U.S. AI Regulation Guide: Legislative Overview and Practical Considerations," *The Journal of Robotics, Artificial Intelligence & Law* 3, no. 1 (2020), https://www.bakermckenzie.com/-/media/files/people/chae-yoon/rail-us-ai-regulation-guide.pdf.
[62] National Conference of State Legislatures, "Legislation Related to Artificial Intelligence," National Conference of State Legislatures, last modified January 17, 2021, https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx.

would apply to commercial entities that earn over 50 million USD per year; hold the data of over one million consumers or consumer devices; or act as data brokers[63] that buy and sell personal information. Companies would be mandated to conduct impact assessments on high-risk automated decision systems with external third parties when "reasonably" possible,[64] including independent auditors and independent technology experts, in order to evaluate the impacts of the algorithm on "accuracy, fairness, bias, discrimination, privacy, and security."[65] If the impact assessments raise concerns, businesses would be required to "reasonably address" the identified issues in a "timely manner."[66] The Algorithmic Accountability Act expressly delineates that it would not preempt any state law, therefore requiring businesses to remain abreast of any state law developments on algorithms and algorithmic bias. The jurisdictional and geographic scope of the bill excludes a private right of action, and does not apply extraterritorially. The measures prescribed by the bill would be enforced by the Federal Trade Commission (FTC) under Section 5 of the Federal Trade Commission Act on deceptive and unfair acts and practices, or via civil suits brought by the affected state's attorney general[67] (the role of the FTC in acting as a data protection agency will be further discussed in Section VII.iii.). While Congress has yet to vote on the bill (the bill was referred to the Subcommittee on Consumer Protection and Commerce,) it may pave the way for further federal legislation regulating AI across industries.

Shortly after the federal Algorithmic Accountability Act was introduced, New Jersey introduced (and later failed) the New Jersey Algorithmic Accountability Act (NJ A.B. 5430) on May 20, 2019. Like its federal counterpart, the bill would have required commercial entities to conduct impact assessments with independent third parties on "high-risk" automated decision

[63] Chae, "U.S. AI Regulation Guide: Legislative Overview and Practical Considerations," 22.
[64] Algorithmic Accountability Act § 3(b)(1)(C) For additional summary information, see Chae, "U.S. AI Regulation Guide: Legislative Overview and Practical Considerations," 21-22.
[65] Algorithmic Accountability Act § 2(2) and §3(b).
[66] Algorithmic Accountability Act § 3(b)(1)(D).
[67] Algorithmic Accountability Act of 2019, note 27 at § 3(d)-(e).

and information systems that involve personally identifiable information regarding race, political opinion, and religion, among other factors.[68] The bill would also have required companies to record any bias or threats to the security of consumer's personally identifiable information (for example a phone number, social security number, or biometric characteristic) as determined by the impact assessment. New Jersey bill NJ S.B. 1943 (2020), which is currently pending and has been referred to the Senate Commerce Committee, further prohibits discrimination by automated decision systems in financial services, insurance, and healthcare services.

IV.ii. Government agency algorithms

Other state and city governments are moving toward the regulation of algorithmic bias in the context of AI procurement and use by government agencies. New York City enacted the first US algorithm accountability law in early 2018 (Int. No. 1696-2017) establishing a task force that would advise on how agencies should inform the public of their use of algorithms, and how agencies should address harms caused by agency algorithms.[69] Washington State introduced and later enacted bills WA S.B. 5527 and WA H.B. 1655 in January of 2019 with prohibitions against algorithmic discriminations, likewise limited to the government's procurement and use of these algorithms (discussed in greater detail below).[70]

IV.iii. Washington AI bias legislation

Washington enacted bill WA H.B. 1655 in 2019, mandating public agencies that develop, procure, or use an automated decision system to complete an algorithmic

[68] New Jersey Algorithmic Accountability Act § 3. For additional summary information, see Chae, "U.S. AI Regulation Guide: Legislative Overview and Practical Considerations," 22-23.
[69] For additional summary information, see Chae, "U.S. AI Regulation Guide: Legislative Overview and Practical Considerations," 23.
[70] Ibid.

accountability report. Public agencies are required to provide notice to individuals impacted by an algorithm and inform these individuals of the following: the automated decision system's name, vendor, and version; what decisions the system will make or support; whether the algorithmic decisions are final or in support of a human-made decision; what policies and guidelines apply to the algorithm's use; and how an affected individual can contest any decision made involving the system. The agency must ensure the automated decision system and the data used by the system are made freely available by the vendor before, during, and after deployment. This is to allow for agency or independent third-party testing, auditing, or research to evaluate the algorithm's potential biases, inaccuracies, or disparate impacts. The agency must also ensure that any decision made or informed by the automated decision system is subject to appeal and immediate suspension if a legal right, duty, or privilege is compromised by the decision. The agency must allow for potential reversal by a human decisionmaker through a timely process accessible to individuals impacted by the algorithm's decision. The agency or vendor must also explain the basis for the automated decision system's output in terms understandable to a layperson.

IV.iv. California AI bias legislation

California introduced and failed bill CA S.B. 444 in February of 2019, requiring businesses that use AI to deliver a product to a public entity to disclose measures taken to reduce "bias inherent in the artificial intelligence system."[71] California bill CA A.B. 2269 failed to enact the Automated Decision Systems Accountability Act of 2020, requiring businesses in California using automated decision systems to proactively and continually test for biases

[71] CA S.B. 444 § 3 For additional summary information, see Chae, "U.S. AI Regulation Guide: Legislative Overview and Practical Considerations," 23.

during the development and deployment of automated decision systems.[72] This testing would be based on impact assessments that would determine whether a protected class was disproportionately, adversely impacted by the algorithm.

IV.v. Illinois AI bias legislation

In 2019, the Illinois General Assembly enacted the Artificial Intelligence Video Interview Act (IL H.B. 2557), requiring employers to notify applicants being interviewed by an AI system that AI may be used to analyze their interview and consider the applicant's fitness for the position. Employers are mandated to provide applicants with information detailing how the algorithm functions and what types of characteristics the algorithm uses to evaluate applicants.[73] Employers need to obtain consent from the applicant to be evaluated by the AI system, and are not permitted to share applicant's videos unnecessarily, deleting an applicant's interview if requested to do so by the applicant.[74] In 2020, bill IL H.B. 4977 failed, attempting to amend the aforementioned Artificial Intelligence Video Interview Act. The amendment to the bill attempted to require employers to gather and report certain demographic information to the Department of Commerce and Economic Opportunity, after which the department would analyze this data and report to the Governor and General Assembly on whether the data suggests a racial bias in companies' hiring algorithms.[75]

IV.vi. Facial Recognition Technology

Within Facial Recognition Technology (FRT), the Commercial Facial Recognition Privacy Act (S. 847) was introduced in March of 2019 and has since been referred to the

[72] Automated Decision Systems Accountability Act Sec. 2. 1798.402. (a).
[73] Artificial Intelligence Video Interview Act § 5(1)-(3). For additional summary information, see Chae, "U.S. AI Regulation Guide: Legislative Overview and Practical Considerations,"28.
[74] Artificial Intelligence Video Interview Act § 10 and § 15.
[75] IL H.B. 4977, Amendment to the Artificial Intelligence Video Interview Act §11-23.

Committee on Commerce, Science, and Transportation. This Act aims to strengthen consumer protections and increase transparency with respect to bias in FRT by requiring companies to conduct meaningful human review of the output of FRT, evaluating whether the output could result in a reasonably foreseeable harm or be "highly offensive" to a reasonable end user.[76] If the algorithm is available as an online service, the company would additionally be required to provide an Application Programming Interface (an interface that allows two applications to talk to each other) to enable third parties to conduct independent tests for accuracy and bias.[77]

IV.vii. Federal AI bias legislation

Within research and development, federal bills H.R. 2202 on the Growing Artificial Intelligence Through Research Act (GrAITR) and S. 1558 on the Artificial Intelligence Initiative Act (AI-IA) direct the President to establish a National AI Research and Development Initiative, strengthening AI research and development by "identifying and minimizing inappropriate bias in datasets, algorithms, and other aspects of artificial intelligence."[78] These bills were introduced in April and May of 2019, and were referred to the House Committee on Science, Space, and Technology, and the Committee on Commerce, Science, and Transportation respectively. The bills also seek to support interdisciplinary research on the societal and ethical implications of AI in order to limit "inappropriate bias" in training data through the establishment of a research and education program on AI and AI engineering.[79] Federal bills H.R. 2575 and S. 1363 on the AI in Government Act were introduced in May of 2019 and placed on the Senate Legislative Calendar (No. 531 and No. 456 respectively). These bills seek to establish an AI Center of Excellence, whose responsibilities

[76] Commercial Facial Recognition Privacy Act § 3(c)(1)-(2). For additional summary information, see Chae, "U.S. AI Regulation Guide: Legislative Overview and Practical Considerations,"23-25.
[77] Commercial Facial Recognition Privacy Act § 3(d).
[78] Growing Artificial Intelligence Through Research Act § 101(6)(F). For additional summary information, see Chae, "U.S. AI Regulation Guide: Legislative Overview and Practical Considerations," 29.
[79] Ibid.

would include studying the economic, legal, ethical, and policy challenges to the use of AI by the federal government and establishing best practices for identifying, assessing, and mitigating biases.[80]

While these bills may pave the way for procedural regularity on harms caused by biases in AI, the great majority of federal and state regulation failed to be enacted. Furthermore, the bills that were introduced provided little guidance on how to identify and mitigate algorithmic biases. This allows for significant leeway and may stall accountability for the biased outcomes of AI.

V. Cases

Within predictive policing, healthcare, education, employment, and other sectors, the implementation of automated algorithmic systems has resulted in numerous legal challenges. Several cases in state and federal jurisdictions address harms caused by algorithmic biases across different industries. Courts have largely been permissive of the continued use of algorithms.

V.i. Predictive policing

Courts have recognized the inconsistent recommendations provided by judicial algorithms, but have continued to endorse their application "if used properly" (*State v. Loomis*). The following cases speak to how courts have considered the application of judicial algorithms for the sentencing of convicted offenders (also referred to as actuarial risk assessment tools). A ProPublica study of COMPAS, a U.S. actuarial risk assessment tool, found that COMPAS was "particularly likely to falsely flag Black defendants as future criminals, wrongly labeling them

[80] Artificial Intelligence Initiative Act § 3(b)(5) and 4(a)(3). For additional summary information, see Chae, "U.S. AI Regulation Guide: Legislative Overview and Practical Considerations," 29.

this way at almost twice the rate as White defendants."[81] The study further indicated that the algorithm's risk assessment score proved highly unreliable in predicting violent crime: only 20% of those predicted to commit violent crimes went on to do so.[82] ProPublica concluded that "the algorithm was somewhat more accurate than a coin flip."[83] Minorities historically have been, and continue to be discriminated against. This bias further perpetuates their second-class status by subjecting them to longer periods of incarceration.

In *State v. Loomis* (2016), the defendant alleged that the circuit court's use of a COMPAS risk assessment (the aforementioned sentencing algorithm) violated his right to due process on the following grounds: first, the algorithm violated his right to be sentenced based on accurate information, in part because the proprietary nature of the algorithm prevented Loomis from being able to evaluate its accuracy; second, the algorithm violated his right to an individualized sentence; and third, the algorithm improperly used a gendered assessment in its sentencing.[84] The court recognized that the proprietary nature of COMPAS prevented the disclosure of the factors weighed by the algorithm and how the algorithm determined risk scores.[85] The court explicitly noted that the algorithm compared defendants to a national sample, but was not tested against the Wisconsin population (where it was applied).[86] The court also recognized that some studies of the COMPAS risk assessment scores have found that these scores disproportionately classify minority offenders as having a higher risk of recidivism.[87] Nevertheless, the court held that providing information on the limitations of the algorithm will enable sentencing courts to better assess the accuracy of its recommendation and how much weight should be given to its risk score. The court concluded: "if used properly, observing the

[81] Julia Angwin et al., "Machine Bias," *ProPublica*, May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[82] Ibid.
[83] Ibid.
[84] *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016) § 34.
[85] Ibid., § 51.
[86] Ibid., § 66.
[87] Ibid., § 59-63.

limitations and cautions set forth herein, a circuit court's consideration of a COMPAS risk assessment at sentencing does not violate a defendant's right to due process."[88]

In *State v. Guise* (2018), the Iowa Court of Appeals evaluated a sentencing algorithm in Iowa – the Iowa Risk Revised (IRR) – and noted that the trial court had "no information on what the IRR was intended to measure, how it was scored, what factors were considered in arriving at a score, or how the PSI evaluator applied the test to Guise."[89] The court also stated that "the broad general language of Iowa Code §§ 901.2(1), 901.5, and 901.3(1)(a) cannot be read to authorize the use of an unspecified algorithm in sentencing. Even courts that have approved the use of algorithms at sentencing have set parameters for their use."[90] Nonetheless, this court and later Iowa's Supreme Court held that the results of certain sentencing algorithms are appropriate for judicial consideration at sentencing, and if used properly, a circuit court's consideration of a risk assessment at sentencing does not violate a defendant's right to due process (as in *State v. Loomis).*

In *State v. Gordon* (2018), the court also found that the use of a sentencing algorithm was permissible. The defendant's risk assessment score was derived from a psychosexual evaluation including two sex-offender risk assessment tools: STATIC-99R and SOTIPS. Gordon's STATIC-99R score indicated he was a level III, average risk for recidivism, while his SOTIPS score indicated that he was at high risk of recidivism.[91] The Court of Appeal found that there was no statutory authority for using these scores for sentencing purposes. This court also noted: "the fact that the algorithm calculates scores based on group data effectively shoehorns a defendant into a grouping score."[92] Despite the inconsistent findings of the two

[88] Ibid., § 104.
[89] *State v. Guise*, 919 N.W.2d 635 (Iowa Ct. App. 2018).
[90] Ibid.
[91] *State v. Gordon*, 919 N.W.2d 635 (Iowa Ct. App. 2018).
[92] Ibid.

risk assessment tools, the Supreme Court of Iowa overturned the Court of Appeal and found that the district court had a right to rely on the risk assessments.

From a regulatory perspective, the federal Justice in Forensic Algorithms Act (H.R. 4368) introduced in September of 2019 and referred to the Subcommittee on Courts, Intellectual Property, and the Internet, would prohibit companies from withholding information on trade-secrecy grounds from a defendant in a criminal proceeding about their automated decision systems, such as the algorithm's source code.[93]

While case law is developing with respect to judicial algorithms and their application to predictive policing, no common legal standard has been established to govern courts' reliance on risk assessment tools.

V.ii. Education

Within education, courts have began relying on standards of explainability – whether the results or recommendations of an algorithm are understandable to lay people – to ensure accountability for harms caused by biases in AI.

Richardson v. Lamar County Bd. of Education (1989) addressed discrimination caused by a teacher certification exam that was processed by an algorithm. The plaintiff's ex-employer claimed that Richardson's employment contract had not been renewed because she failed to pass the certification exam.[94] The plaintiff alleged that her contract was not renewed either because of her race (under disparate treament), or because the exam had a disparate impact on herself and fellow African-American teachers.[95] The court sided with the plaintiff on her

[93] AI Now, "AI Now 2019 Report," 34.
[94] *Richardson v. Lamar County Bd. of Education*, 729 F. Supp. 806 (U.S. Dist 1989)
[95] Ibid.

disparate impact claim largely on the contents of the exam rather than algorithmic harm, but did not uphold her disparate treatment claim.[96]

Houston Federation of Teachers v. Houston Independent School District (2017) similarly challenges the use of an algorithm processing standardized assessments of teachers.[97] The Educational Value-Added Assessment System (EVAAS) was employed by the Houston Independent School District to improve teaching quality through the automated evaluation of these assessments. The court ruled in favor of the plaintiffs on procedural due process grounds, recognizing teachers' property interest in their continued employment, and the teachers' inability to access, understand, or act on the algorithm's findings.[98]

The latter case exemplifies courts' increasing reliance on explainability. Explainability is developing into a standard of evidence for algorithmic harms, despite not being central to the identification or mitigation of harm (discussed in greater detail in Part VIII.ii.).

V.iii. Employment

Courts have largely ruled against plaintiffs alleging harm in cases addressing employment algorithms. In *Coleman v. Exxon Chem. Corp.* (2001), plaintiffs alleged that an employee ranking system, administered by an algorithm which rated employees and determined their salaries, allowed for "racial and/or gender biases of supervisor-rankers to run unchecked."[99] The court ruled against the plaintiffs on the grounds that the ranking system was facially neutral, and the employees were unable to prove pretext or unlawful intent, in addition to their evidence being deemed inadmissible.[100] Other employment claims for racial

[96] Ibid.
[97] *Houston Federation of Teachers v. Houston Independent School District*, 51 F. Supp. 3d 1168 (S.D. Tex. 2017)
[98] Ibid.
[99] *Coleman v. Exxon Chem. Corp.*, 162 F. Supp. 2d 593 (U.S. Dist 2001)
[100] Ibid.

discrimination involving algorithms have also failed (*Dawson v. Phila. Media Holdings*; *Damino v. City of New York*).

Bauserman v. Unemployment Ins. Agency (2019) addresses the application of the Michigan Integrated Data Automated System (MiDAS) algorithm by the Michigan Unemployment Insurance Agency to investigate benefits fraud and penalize those who allegedly commit it.[101] Affected individuals were sent prepopulated online questionnaires that triggered automatic findings of fraud in many cases.[102] Automatic determinations of fraud also occurred if recipients failed to respond within ten days, or when the MiDAS algorithm deemed their responses unsatisfactory.[103] The algorithm automatically categorized any discrepancies as fraud, falsely accusing more than 40,000 people of fraud.[104] The plaintiffs experienced devastating consequences, including tax-refund seizures, wage garnishment, and civil penalties without notice. The Michigan Supreme Court sided with the plaintiffs.[105]

Bauserman v. Unemployment Ins. Agency highlights how far reaching the consequences of biases in algorithms employed by state agencies can be in compromising the financial stability of several tens of thousands of individuals.

V.iv. Healthcare

Within healthcare, courts have demonstrated varied approaches to mitigating harms caused by biases in healthcare algorithms, whether through the disclosure of source code, the mandated creation of a new algorithm, or the reinstatement of benefits.

K.W. v. Armstrong (2014) demonstrated how an algorithm used by Idaho's state Medicaid program to determine Medicaid payments for adults with intellectual and

[101] *Bauserman v. Unemployment Ins. Agency*, 503 Mich. 169 (2019).

[102] Ibid.
[103] Ibid.
[104] Ibid.
[105] Ibid.

developmental disabilities resulted in drastic drops in participants' payments, leading to horrific living conditions for those who no longer received sufficient hours of in-home care and services.[106] The plaintiffs filed a class action lawsuit which was settled. Preliminarily, the court ordered the state to disclose its formula, fix the formula so that participants received the proper amount of funds, and develop and implement procedural protections for those who had already been impacted.[107] As part of the settlement, the state agreed to develop a new formula and provide participants with the highest dollar amount of payments possible under the existing algorithm's recommendations, until the new formula was implemented.[108]

Ark. Dep't of Human Servs. v. Ledgerwood (2017) similarly addressed a Medicaid algorithm that drastically reduced the Medicaid attendant care hours for many low-income adult Medicaid participants living with disabilities in Arkansas.[109] Consequently, many participants experienced terrible living conditions. After the court ordered an injunction against the use of the algorithm, the Department of Human Services issued an emergency rule and began applying the same algorithm for two more months despite the injunction.[110] The DHS thereafter employed a similar automated decision making system, while allowing expert nurses to conduct individualized assessments and use their discretion for the number of attendant care hours patients receive.[111]

While not in healthcare, *Barry v. Lyon* (2016) addressed the use of a matching algorithm by the Michigan Department of Health and Human Services (MDHHS) to disqualify individuals for food assistance if they were determined to have an outstanding felony warrant.[112] Over 19,000 people were improperly disqualified and given only vague notice of

[106] *K.W. ex rel. D.W. v. Armstrong*, 298 F.R.D. 479 (D. Idaho 2014).
[107] Ibid.
[108] Ibid.
[109] *Ark. Dep't of Human Servs. v. Ledgerwood*, 530 S.W.3d 336 (Ark. 2017)

[110] Ibid.
[111] Ibid.
[112] *Barry v. Lyon*, 834 F.3d 706 (6th Cir. 2016)

their disqualification.[113] The district court ruled that the automatic disqualification policy violated the federal Supplemental Nutrition Assistance Program (SNAP), the constitutional Supremacy Clause, and constitutional and statutory due process.[114] The district court ruled that people's benefits had to be reinstated, and the Sixth Circuit upheld this ruling.

K.W. v. Armstrong offers an innovative approach to recognizing and meeting the needs of plaintiffs while allowing time for Idaho's state Medicaid program to retrain their algorithm.

These cases across form a patchwork of legal perspectives on how courts have evaluated algorithmic biases in different sectors, and the lack of a foundational precedent or unified legal approach to these cases informs the need for a theory of algorithmic liability. Notably, most courts ruled against plaintiffs alleging algorithmic harm. This further challenges the process of extrapolating a theory of algorithmic liability from existing case law.

VI. Need for a theory of algorithmic liability

The lack of clear and consistent guidance on algorithmic liability in regulation and case law motivates the development of an algorithmic duty of care in order to hold all entities involved in the development and deployment of an algorithm accountable to anticipating and mitigating algorithmic harms.

VI.i. Duty of technology competence

The duty of technology competence instated by the American Bar Association in 2012 requires lawyers to keep abreast of "changes in the law and its practice, including the benefits

[113] Ibid.
[114] Ibid.

and risks associated with relevant technology."[115] Baker (2017) interpreted that the language of this duty was left purposefully broad to account for the technology of today, as well as technology that has not yet been conceived.[116] While this duty does not explicitly include algorithms, it can be assumed that AI falls within the bounds of the duty of technology competence. In guidance issued by the California Bar Association, lawyers were given three approaches to handling unfamiliar technology: "(1) become familiar with the technology, (2) consult with or delegate to someone who is familiar with the technology, or (3) decline to represent the client."[117] This motivates the need for a theory of algorithmic liability for harms caused by biases in AI to ensure competent representation.

VI.ii. Duty of care for platforms

The technological duty of care has only been theorized with respect to platforms such as social media companies. Platforms have a duty of care under Section 230 of the Communications Decency Act[118] to moderate content through notice and takedown. This duty is not only retroactive, but also proactive, ensuring that harmful content does not have the opportunity to be uploaded and shared. However, platforms are only required to tackle illegal content – they are only required to take down content that is prohibited by law, not content that might broadly be deemed 'harmful,' such as hate speech. Furthermore, Section 230 requires what could be likened to a good samaritan approach, because platforms are only required to take down illegal content that they have come across. Consequently, platforms may be

[115] American Bar Association, "Rule 1.1 Competence - Comment," American Bar Association, last modified 2021, https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_1_competence/comment_on_rule_1_1/.
[116] Jamie J. Baker, "Beyond the Information Age: The Duty of Technology Competence in the Algorithmic Society," South Carolina Law Review 69 (2018): 557, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3097250#:~:text=Jamie%20Baker,-Texas%20Tech%20University&text=As%20society%20moves%20beyond%20the,in%20this%20brave%20new%20world.
[117] Ibid., 566.
[118] 47 U.S. Code § 230.

incentivized *not* to monitor content – to limit their knowledge and control of content – as they will not be held liable for content they did not know of. This duty of care for platforms does not translate appropriately into an algorithmic context, which further motivates the need for a duty of care for algorithms.

VI.iii. Tort law

United States v. Caroll Towing Co. (1947)[119] a foundational precedent to modern American tort law, will become the backbone of the proposed duty of care to ensure accountability for biases in algorithmic systems. This case resolved a dispute regarding damage to a barge and resultant lost cargo. The defendant's barge carried a cargo of flour owned by the United States. After the mooring lines of the barge were adjusted at the pier, the barge broke free and hit a tanker, causing the barge to dump its cargo and sink. No one was on board of the barge at the time of these events. The question before the court was whether the defendant could be held liable for the damage to the barge and the lost cargo by not having anyone aboard the barge when the barge broke free. The court held that the defendant was partly liable as the burden of having someone aboard the barge was less than the loss inflicted, multiplied by the probability of the barge breaking free when left unattended. In his judgment of the case, Judge Hand proposed the following formula:

$$B < P \cdot L$$

where the burden of care is B; the probability, P; and the loss or injury, L. Liability depends upon whether B is less than P multiplied by L, i.e. whether the burden was less than the harm multiplied by the probability for determining duty (the probability that the plaintiff had a

[119] *United States v. Carroll Towing Co.*, 159 F. 2d 169 (2d. Cir. 1947).

preexisting duty of care). This formula will inform the proposed duty of care for algorithms in balancing the burden of care with the risk of harm.

VII. Proposed duty of care for algorithms

The formula introduced in Section VI will become the foundation of the proposed duty of care for algorithms, adapted in order to best identify parties that may be implicated for algorithmic harms. This formula will serve as a test, applied to each party involved in the development and deployment of an algorithm, to determine if that party reasonably should have known that harm would occur, and if that party engaged in action to prevent such harm.

The proposed algorithmic duty of care is as follows:

$$B - \frac{k}{\min\{A\}} < P \cdot L$$

where B is the burden of care in having the automated task be completed manually by a human; k is a constant determining the influence of A; A is the minimum cost of an alternative algorithmic solution (retraining the existing algorithm, creating a new algorithm, or using an existing algorithm that achieves the same legitimate purpose with a lesser impact on protected classes); P is the probability of a preexisting duty of care as determined by a disparate impact assessment evaluating the extent of any disproportionate, adverse impact on a protected class (identified through auditing of the algorithm); and L is the loss or harm from algorithmic bias experienced by a protected class.

Liability for algorithmic harm would effectively depend on whether the combined burden of care (a human completing the automated task) less the cost of using an alternative algorithmic solution is lower than the probability of the party having a preexisting duty of care multiplied by the algorithmic harm caused. Judges are afforded discretion in deciding how

influential they determine the analysis of alternative algorithmic solutions to be in assigning a value to the constant k. For the purposes of this paper, k will be assumed to be 1.

In practice, A serves as a penalty. If the minimum cost of an alternative algorithmic solution is high, then the defendants are penalized less. Should the minimum cost of an alternative algorithmic solution be low, the defendants would be penalized more. The magnitude of A's penalty is mediated by k. In addition to scaling the cost of alternative algorithmic solutions, k should also account for the accuracy of the algorithmic solutions vis-à-vis their human counterpart. If the accuracy of the human solution is higher than that of the alternative algorithmic solutions, the defendants should be penalized more for employing a biased and inaccurate algorithm. Meanwhile if the human solution is less accurate than the alternative algorithmic solutions, the defendants should be penalized less for choosing not to implement a less accurate human solution. In order to limit the burden of identifying and implementing alternative algorithmic solutions, A must be the minimum cost of retraining the existing algorithm, creating a new algorith, or utilizing an alternative existing algorithm.

Generally, liability will be proportional to the level of automation in the product or service causing harm. The more automation, the more care is required to mitigate potential harms to society, in particular harms to vulnerable and marginalized groups. Higher levels of automation translate into a higher probability of there being a preexisting duty of care (P), increasing the likelihood of a duty of care being established.

VII.i. Jurisdiction

Three jurisdictional issues frame the domain to which this theory applies: first, the nature of the person or entity affected; second, the holder of the duty; and third, the geographic location of the algorithm or the data processed by the algorithm.

In the absence of a geographically bound US data protection regulation, this theory applies only to US citizens, residents, and non-resident aliens affected by an algorithm on US territory, in line with federal and state civil rights statutes.[120] These citizens, residents, or non-resident aliens must be identifiable and must be living, as unidentifiable or deceased individuals do not have personal data (any information relating to an identifiable natural person) from which they could suffer or experience harm.

The lifecycle of data informs who may hold a duty of care: collectors of data, processors of data, analyzers of data, sharers of data, and storers of data may all be involved in the production of algorithmic harm against a protected class. Parties involved in the training of data, the modeling of an algorithmic solution, and the application of an algorithmic tool may all be held liable for the harm caused. Section 230 of the Communications Decency Act, however, may foreclose the ability to prosecute a company for an algorithm published on their platform.

The proposed duty of care applies to both public and private algorithms on the grounds that the harm addressed is not sector specific, and the enforcement of auditing through disparate impact assessments should not differ among public or private algorithms.

The movement of data between jurisdictions, particularly national jurisdictions, complicates the legal standing of a victim to pursue suit. The interjurisdictional nature of data transit among processors (those processing data on behalf of the controller) and controllers (those determining the purpose and means of processing) means that data may not be located in the same place as the location of harm. The geographic scope of this duty of care is limited to the US state in which the output or recommendation of the algorithm harmed the victim, unless the harm was suffered by a collective on a national scale, rendering the case a national class action suit (discussed in Section VII.iv.).

[120] Legal Information Institute, "Alien," Legal Information Institute, last modified 2021, https://www.law.cornell.edu/wex/alien#:~:text=Generally%2C%20both%20legal%20and%20illegal,in%20United%20States%20federal%20court.&text=U.S.%20courts%20typically%20grant%20nonresident,arose%20within%20the%20United%20States.

VII.ii. Auditing of algorithms

The auditing of algorithms relies on operations data and systems data. An operational audit examines the data an algorithm uses to calibrate its function.[121] A systems audit reviews the quality of the function the algorithm is performing.[122] Operations and systems data can overlap, be different presentations of the same dataset, or be distinct datasets. At the very least, a systems audit must be performed in order to complete a disparate impact assessment.

In the case of the Impact Pro algorithm, Obermeyer et al. (2019) were able to conduct both an operational audit and systems audit in reverse engineering the algorithm. The authors determined disparate impact by comparing the expected interest for someone given their race, denoted as either Black or White, and given equal risk scores, as follows:[123]

$$\mathbb{E}(Y|R, W) = \mathbb{E}(Y|R, B)$$

where E is the expectation, and Y is the interest given a risk score R and the race of the patient (W or B). If the conditional expectations were equal, the authors interpreted this an absence of bias.[124] If the conditional expectations significantly differed, the authors took this difference as evidence of bias, and in the case of Impact Pro, racial discrimination.[125]

VII.iii. Federal Trade Commission as data protection agency

Auditing is complicated by the fact that the United States lacks a data protection agency, which could serve as an independent auditor of algorithms. In the absence of a data protection agency, the United States has relied on the Federal Trade Commission (FTC) as a default enforcement mechanism, adopting these responsibilities as part of its duties to act on

[121] S. A. Sayana, "The IS Audit Process," Information Systems Control Journal 1 (2002), http://carl.sandiego.edu/ctu/IS_audit_process.pdf.
[122] Ibid.
[123] Obermeyer et al., "Dissecting racial bias in an algorithm used to manage the health of populations," 448.
[124] Ibid.
[125] Ibid.

deceptive and unfair acts and practices. The FTC is not equipped, neither statutorally nor technically, to take on the responsibilities of a data protection agency. The FTC chose to become de facto responsible for data protection without any delineated statutory authority.

The FTC conducts investigations in order to protect consumers and promote competition, suing companies and individuals that violate the law, developing rules to ensure a vibrant marketplace, and educating consumers and businesses about their rights and responsibilities.[126] In order to adopt the responsibilities of a data protection agency, the FTC should handle reports of data breaches, interpret and enforce the Fair Information Practice Principles (data protection principles that apply to federal agencies) alongside state and federal data protection regulations, and educate businesses on proper data protection protocols. With respect to the proposed duty of care, the FTC should be empowered to impartially and independently conduct disparate impact assessments and enforce fines or other penalties assigned by courts.

This duty strives to address harms of representation, harms that stem from the intentional or unintentional subordination of certain protected classes, as well as harms of allocation, harms that stem from some protected classes being denied, or having restricted access to valuable resources and opportunities.[127] Harms of representation are best exemplified by *State v. Loomis* detailed earlier, in subordinating defendants of minority races and ethnicities and subjecting these to longer sentences due to the misattribution of their recidivism risk. *Ark. Dep't of Human Servs. v. Ledgerwood* and *Barry v. Lyon* represent harms of allocation, in which the defendants were denied access to the valuable resources of attendant care hours and food assistance. Both of these harms are made particularly salient by the misuse of group data (as

[126] Federal Trade Commission, "What We Do," Federal Trade Commission, last modified April 15, 2014, https://www.ftc.gov/about-ftc/what-we-do.
[127] Dillon Reisman et al., "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability," *AI Now*, April 2018, 18, https://ainowinstitute.org/aiareport2018.pdf.

opposed to the data of an individual processed by an algorithm), having the potential to harm entire communities or populations of individuals.

VII.iv. Group data and class action suits

Three considerations frame the application of this theory to algorithmic harms caused by the misuse of group data: first, the challenge of addressing class-specific harm as opposed to individual harm; second, identifying harm to a specific class; and third, assigning causation between a class and a harm. The classes to which the proposed duty of care applies will continue to be protected classes, as these are encoded under federal nondiscrimination statutes. As defined in the introduction, protected class will refer to race, color, religion, national origin, citizenship, sex (including gender, pregnancy, sexual orientation, and gender identity), age, physical or mental disability, familial status, veteran status, and genetic information.

The first two considerations will be addressed by existing legal standards for class action lawsuits. Class action lawsuits remedy physical or financial harm committed against groups of individuals by corporations.[128] Notably, these lawsuits allow groups to take legal action against an entity in circumstances when filing individual lawsuits would either burden courts or be financially unviable.[129] The challenge of addressing class-specific harm will be addressed through a class action framework.

The identification of class-specific harm is rooted in the way that class is certified by judges. The members of a class must be individuals who experienced the *same* injury or harm, in addition to having "factual and legal issues that are common to all class members."[130] The identification of harm specific to a class will similarly hinge on whether all members of a class

[128] ClassAction.org, "How to Start a Class Action Lawsuit," ClassAction.org, last modified 2021, https://www.classaction.org/learn/how-to-start#requirements.
[129] Ibid.
[130] Ibid.

experienced the same injury or harm, and whether these members share the same factual and legal grounds for claiming discrimination under a disparate impact framework.

Determining L – the loss or injury suffered by a group – is difficult, particularly when certain members of that group or no members of that group can be identified. Additionally, specific sub-groups or individuals may be unaware of the harm committed against them. The loss, injury, or harm suffered by a group may also be unevenly distributed. Legal standards of class action will provide some clarity for these circumstances. In the first and second case, class members, while not initially aware of the suit, may be made aware of the suit through a class action notice received by mail, or media coverage of the suit.[131] As in the case of class action lawsuits, class members can join an existing suit by contacting the office representing the named plaintiff and providing documentation (if available) of the algorithm's discriminatory outcome against them. In the third case, should an individual believe they have suffered greater injury than other members of their class, they should pursue their own suit, as damages and reparations are awarded equally among all members of the injured class if the accused party or parties are found at fault.[132]

The application of this theory to algorithmic harms involving group data hinges on the disparate impact assessment determining L for the members of a class. Determining whether a class is harmed for reasons related to being a class, or for reasons unrelated to class allows us to assign causality to harm committed by a biased algorithm.

While a descriptive approach to auditing, as in the case of Obermeyer et al.'s audit of Impact Pro, the technique of propensity score matching may allow for a close approximation of causality between class and harm. Propensity score matching is a technique that attempts to reduce potential bias when estimating the effect of a treatment or characteristic by matching

[131] ClassAction.org, "How to Start a Class Action Lawsuit."
[132] Ibid.

similar observations.[133] This statistical technique would be akin to a systems audit. Alongside the algorithm's output of interest (similar to Obermeyer et al.'s expectation of interest), this technique depends on the collection of a vast range of characteristics about an individual beyond their class (in the case of Impact Pro, their race), such as their age, past medical history, and level of education. A propensity score is calculated based on the data available about each individual affected by the algorithm's recommendation, and the propensity score is used to match pairs of individuals on all other characteristics apart from their class.[134] These pairs of individuals would therefore be identical in age, have very similar past medical history, and a very similar income (among other characteristics). Therefore, we can be confident that the comparison is unbiased by characteristics other than the class considered.

After accumulating many matched pairs of individuals, the difference in algorithmic output of interest between each matched pair is calculated.[135] If a difference is identified between classes and this difference is statistically significant, and the only characteristic distinguishing the members of the pair is their class, then the algorithm is deemed discriminatory or biased. In order to conduct propensity score matching, descriptive data about all individuals affected by the algorithm would need to be requested during discovery, the investigatory phase of a lawsuit.

While this statistical method is descriptive in nature, it strongly implies a causal link where causation between class and harm can never be definitively proven. Without a randomized experiment with a control group and treatment group, a causal link cannot be proven. Recognizing the infeasibility of conducting randomized experiments for each class action lawsuit, statistical inference methods based on past data are best suited for inferring a causal relationship between class and harm.

---

[133] Peter C. Austin, "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate Behavioral Research* 46, no. 3 (2011): 399, doi:10.1080/00273171.2011.568786.
[134] Ibid., 405.
[135] Ibid.

As with most statistical methods, propensity score matching may be limited by other confounding variables that are not included in the descriptive dataset of individuals affected by the algorithm. For example, income is often times a confounding variable that may be correlated to the variable of interest, such as a referral for a higher degree of specialized care, in the case of Impact Pro.

VII.v. Sustained development of proposed duty of care

To facilitate the sustained development of the proposed duty of care for algorithms, the auditing of datasets, models, and training techniques must become commonplace. Rigorous testing should be required across the lifecycle of AI systems through pre-release trials (the testing of algorithms prior to their deployment), as well as ongoing audits over the course of an algorithm's runtime.[136] Establishing a legal environment in which records of data processing activities are maintained, as mandated by the General Data Protection Regulation (GDPR) in Europe, empowers users to challenge the data, variables, and output of an algorithm through independent auditing.

Within this environment, it is important to consider the levels of technological literacy[137] of users, regulators, and other involved parties. This requires weighing the burden for users to seek out information about whether an algorithm is processing their personal data versus the burden for companies to inform users that their data is being processed by an algorithm. In the early development of motorvehicles, consumers were not expected to install seatbelts in their cars if they wished to do so; car manufacturers who were mandated to do so

[136] West, Whittaker, and Crawford, *Discriminating Systems: Gender, Race, and Power in AI*, 4.
[137] An individual's ability to use and understand technology to assess, acquire and communicate information.

by regulators.[138] If the burden of technological literacy is not considered, the digital underclass[139] will be placed most at risk.

VIII. Conclusion and limitations

The proposed theory of algorithmic liability seeks to address discriminatory harms caused by the first generation of AI - algorithms that are reactive to input data based on a program or model. Informed by the foundational precedent of tort law, *United States v. Caroll Towing Co.*, liability would depend on whether the combined burden of care (a human completing the automated task) less the cost of using an alternative algorithmic solution is lower than the probability of the party having a preexisting duty of care multiplied by the algorithmic harm caused. This theory is jurisdictionally limited by the nature of the victim, the holder of the duty, and the geographic location of the harm. As such, it will apply to identifiable and living US citizens, residents, and non-resident aliens affected on US territory; collectors, processors, analyzers, sharers, and storers of data involved in the development of an algorithm, as well as the companies, organizations, or agencies involved in the deployment of an algorithm; the US state in which the algorithmic output or recommendation harmed the victim; and both public and private algorithms.

VIII.i. Limitations of proposed duty of care for algorithms

The proposed duty of care for algorithms is limited on numerous fronts. First, in targeting discriminatory harms committed against a protected class, it is limited to harms stemming from the misuse of personal information (PI), or personally identifiable information

[138] The installment of seat belts in vehicles has been mandatory since 1968 under the Federal Motor Vehicle Safety Standard 208.
[139] Vulnerable populations who do not use the Internet frequently.

(PII)[140] pertaining to a protected class, such as street address, photographic images, biometric data, social security number, or passport number. Race or religion, while a personal characteristic, does not constitute PI or PII as these traits are shared by more than one person.[141]

This theory may not address harms stemming from action-based information (ABI)[142] – location-based information such as GPS data – and only addresses harms stemming from demographically identifiable data (DII)[143] when it pertains to a protected class. Harms may arise that are intrinsic to these ABI and DII datatypes that are not related to the algorithm, but are related to the typology of data used to train the algorithm. Issues related to these datatypes may need to be addressed specifically, rather than assigning liability to an algorithmic system as a whole.

To reiterate, the proposed theory of algorithmic liability is limited by the nature of the victim, the holder of the duty, and the geographic location of the harm. This duty of care does not apply to unidentifiable or deceased individuals, foreign citizens in the US for temporary purposes (such as tourism or business), algorithmic harm suffered outside of US territories, with victims suing in the state in which they suffered harm (regardless of where their data or the algorithm inflicting the harm is located.)

The Impact Pro algorithm illustrates the challenges we will continue to face in ensuring accountability for algorithmic harms. Litigating the racially discriminatory harm caused by Impact Pro is complicated by the algorithm's reliance on group data, the fact that its subjects

---

[140] Daniel J. Solove and Paul M. Schwartz, "The PII Problem: Privacy and a New Concept of Personally Identifiable Information," *New York University Law Review* 86 (2011), https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=2089&context=faculty_publications.

[141] University of Pittsburgh, "Guide to Identifying Personally Identifiable Information (PII)," Information Technology, last modified February 16, 2021, https://www.technology.pitt.edu/help-desk/how-to-documents/guide-identifying-personally-identifiable-information-pii#:~:text=Personal%20identification%20numbers%3A%20social%20security,Personal%20telephone%20numbers.

[142] Patrick Biltgen and Stephen Ryan, *Activity-Based Intelligence: Principles and Applications* (Artech House, 2016).

[143] Nathaniel A. Raymond, "Beyond 'Do No Harm' and Individual Consent: Reckoning with the Emerging Ethical Challenges of Civil Society's Use of Data," in: Taylor L., Floridi L., van der Sloot B. (eds) Group Privacy, Philosophical Studies Series, vol 126 (2017), https://doi.org/10.1007/978-3-319-46608-8_4.

are largely indiscernible, and the ecosystem in which that the primary users of the algorithm (doctors) are not the victims affected by its biases (patients.)

VIII.ii. Current emphasis on algorithmic explainability

Existing case law and regulation has targeted the 'explainability' of algorithms – whether the results or recommendations of an algorithm are understandable to lay people (including *Houston Federation of Teachers v. Houston Independent School District* and bill WA H.B. 1655). Explainability has become a standard of evidence for algorithmic harms, and algorithmic regulation is stalling in the absence of explainability. In addition to often times being unattainable due to confidentiality surrounding the proprietary code of an algorithm, explainability does not address cases in which subjects do not know their data is being processed, or cases in which the subjects of an algorithm aren't its primary users, as in the Impact Pro algorithm. Explainability mandates an understanding of the algorithm's training data, model, and code, all of which are often compiled by different individuals, companies, or organizations. Explainability is a very high standard for evidence of bias, which is failing us in being a precondition to the regulation and litigation of algorithmic harm. The concept of explainability erodes at our ability to establish stare decisis and develop case law delineating algorithmic liability.

VIII.iii. Measures of success for the litigation of algorithmic harms

The success of litigation in the field of harms caused by biases in AI can be assessed by evaluating whether a foundational precedent has been set with respect to algorithmic harms. We have not had a *Brown v. Board of Education* for algorithms and we have suffered from this lack of guidance. Following the establishment of a super-precedent, structural change within government agencies and programs can be considered a subsequent measure of success.

Overemphasis on the explainability of algorithms and the difficulty of attributing causality to algorithmic harms has placed us in an adjudication gap.

The proposed duty of care and all theories of liability for algorithmic harms will need to evolve as new datatypes are introduced and algorithms develop into the second, third, and fourth generations of AI, introducing novel forms of harm. Tort law can serve to protect vulnerable and marginalized populations from algorithmic harm. In the absence of a duty of care for algorithms, the regulation of harms caused by biases in AI may continue to stall. A duty of care for algorithms provides a mechanism by which judges and legislators can ensure accountability for harms caused by algorithmic biases: a standard to which actors involved in the development and deployment of algorithms will be held to under the law.

IX. Works Cited

AI Now. "AI Now 2019 Report." *AI Now*, 2019, 1-100. https://ainowinstitute.org/AI_Now_2019_Report.pdf.

Akhtar, Allana. "New York is Investigating UnitedHealth's Use of a Medical Algorithm That Steered Black Patients Away from Getting Higher-quality Care." Business Insider. Last modified October 28, 2019. https://www.businessinsider.com/an-algorithm-treatment-to-white-patients-over-sicker-black-ones-2019-10.

American Bar Association. "Rule 1.1 Competence - Comment." American Bar Association. Last modified 2021. https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_1_competence/comment_on_rule_1_1/.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica*, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Ark. Dep't of Human Servs. v. Ledgerwood, 530 S.W.3d 336 (Ark. 2017).

Austin, Peter C. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research* 46, no. 3 (2011), 399-424. doi:10.1080/00273171.2011.568786.

Baker, Jamie J. "Beyond the Information Age: The Duty of Technology Competence in the Algorithmic Society." *South Carolina Law Review* 69 (2018), 557-577. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3097250#:~:text=Jamie%20Baker,-Texas%20Tech%20University&text=As%20society%20moves%20beyond%20the,in%20this%20brave%20new%20world.

Barry v. Lyon, 834 F.3d 706 (6th Cir. 2016).

Bauserman v. Unemployment Ins. Agency, 503 Mich. 169 (2019).

Biltgen, Patrick, and Stephen Ryan. *Activity-Based Intelligence: Principles and Applications.* Artech House, 2016.

Bornstein, Stephanie. "Antidiscriminatory Algorithms." *Alabama Law Review* 70 (2018), 519-572. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3307893.

Buchanan, Bruce G. "A (Very) Brief History of Artificial Intelligence." *AI Magazine* 26, no. 4 (2005), 53-60. https://ojs.aaai.org/index.php/aimagazine/article/view/1848.

CA A.B. 2269

CA S.B. 444

Centers for Disease Control and Prevention. "Tuskegee Study." CDC. Last modified July 16, 2020. https://www.cdc.gov/tuskegee/timeline.htm.

Chae, Yoon. "U.S. AI Regulation Guide: Legislative Overview and Practical Considerations." *The Journal of Robotics, Artificial Intelligence & Law* 3, no. 1 (2020), 17-40. https://www.bakermckenzie.com/-/media/files/people/chae-yoon/rail-us-ai-regulation-guide.pdf.

Chagal-Feferkorn, Karni. "The Reasonable Algorithm." *Journal of Law, Technology and Policy*, 2018. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3095436.

ClassAction.org. "How to Start a Class Action Lawsuit." ClassAction.org. Last modified 2021. https://www.classaction.org/learn/how-to-start#requirements.

Coleman v. Exxon Chem. Corp., 162 F. Supp. 2d 593 (U.S. Dist 2001).

Congress. "H.R.2231 - Algorithmic Accountability Act of 2019." Congress. Last modified 2019. https://www.congress.gov/bill/116th-congress/house-bill/2231#:~:text=Introduced%20in%20House%20(04%2F10%2F2019)&text=This%20bill%20requires%20specified%20commercial,artificial%20intelligence%20or%20machine%20learning.

Cowger, Jr., Alfred R. "Liability Considerations When Autonomous Vehicles Choose The Accident Victim." *Journal of High Technology Law* 14 (2018), 1-60. https://cpb-us-e1.wpmucdn.com/sites.suffolk.edu/dist/5/1153/files/2018/12/Cowger-FINAL-174f0gc.pdf.

Cox, Michael T. "Perpetual Self-Aware Cognitive Agents." *AI Magazine* 28, no. 1 (2007), 32-46. https://doi.org/10.1609/aimag.v28i1.2027.

Cuzzolin, F., A. Morelli, B. Cîrstea, and B. J. Sahakian. "Knowing me, knowing you: theory of mind in AI." *Psychological Medicine* 50, no. 7 (2020), 1057-1061. doi:10.1017/s0033291720000835.

Department of Financial Services. *Letter to CEO of UnitedHealth.* New York: Department of Financial Services, 2019. https://dfs.ny.gov/system/files/documents/2019/10/20191025160637.pdf.

Diakopoulos, Nicholas. "Accountability in Algorithmic Decision Making." *Communications of the ACM* 59, no. 2 (2016), 56-62. doi:10.1145/2844110.

Evans, Melanie, and Anna W. Mathews. "New York Regulator Probes UnitedHealth Algorithm for Racial Bias." Wall Street Journal. Last modified October 26, 2019.

https://www.wsj.com/articles/new-york-regulator-probes-unitedhealth-algorithm-for-racial-bias-11572087601.

Federal Trade Commission. "What We Do." Federal Trade Commission. Last modified April 15, 2014. https://www.ftc.gov/about-ftc/what-we-do.

Georgeff, Michael P., and Amy L. Lanksy. "Reactive Reasoning and Planning." *Robotics*, 1987, 677-682. https://www.aaai.org/Papers/AAAI/1987/AAAI87-121.pdf.

H.R. 2202

H.R. 2231

H.R. 2575

Houston Federation of Teachers v. Houston Independent School District, 51 F. Supp. 3d 1168 (S.D. Tex. 2017).

IBM. *Everyday Ethics for Artificial Intelligence.* IBM, 2019. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

IL H.B. 2557

IL H.B. 4977

K.W. ex rel. D.W. v. Armstrong, 298 F.R.D. 479 (D. Idaho 2014).

Kroll, Joshua A., Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. "Accountable Algorithms." *University of Pennsylvania Law Review* 165 (2017), 633-705. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2765268.

Legal Information Institute. "Alien." Legal Information Institute. Last modified 2021. https://www.law.cornell.edu/wex/alien#:~:text=Generally%2C%20both%20legal%20and%20illegal,in%20United%20States%20federal%20court.&text=U.S.%20courts%20typically%20grant%20nonresident,arose%20within%20the%20United%20States.

Lima, Dafni. "Could AI Agents Be Held Criminally Liable? Artificial Intelligence and the Challenges for Criminal Law." *South Carolina Law Review* 69 (2018), 677-696. https://www.researchgate.net/publication/335107356_Could_AI_Agents_Be_Held_Criminally_Liable_Artificial_Intelligence_and_the_Challenges_for_Criminal_Law.

Mayson, Sandra G. "Bias In, Bias Out." *The Yale Law Journal* 128 (2019), 2218-2300. https://www.yalelawjournal.org/article/bias-in-bias-out.

Myers West, Sarah, Meredith Whittaker, and Kate Crawford. *Discriminating Systems: Gender, Race, and Power in AI.* AI Now, 2019. https://ainowinstitute.org/discriminatingsystems.pdf.

National Conference of State Legislatures. "Legislation Related to Artificial Intelligence." National Conference of State Legislatures. Last modified January 17, 2021. https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx.

NJ A.B. 5430

NJ S.B. 1943

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366 (2019), 447-453. doi:10.1126/science.aax2342.

OPTUM. "Impact Pro: Individual & Population Health Risk Analytics." OPTUM. Last modified 2021. https://www.optum.com/business/solutions/data-analytics/data-analytics-health-plans/impact-pro-cpl.html.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability." *AI Now*, April 2018, 1-22. https://ainowinstitute.org/aiareport2018.pdf.

Richardson v. Lamar County Bd. of Education, 729 F. Supp. 806 (U.S. Dist 1989).

Richardson, Rashida, Jason M. Schultz, and Vincent M. Southerland. "Litigating Algorithms 2019 US Report." *AI Now*, 2019, 1-32. https://ainowinstitute.org/litigatingalgorithms-2019-us.pdf.

S. 1108

S. 1363

S. 1558

S. 847

Sayana, S. A. "The IS Audit Process." *Information Systems Control Journal* 1 (2002). http://carl.sandiego.edu/ctu/IS_audit_process.pdf.

Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department. "Online Harms White Paper." British Parliament. Last modified 2019. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf.

Solove, Daniel J., and Paul M. Schwartz. "The PII Problem: Privacy and a New Concept of Personally Identifiable Information." *New York University Law Review* 86 (2011), 1814-1894. https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=2089&context=faculty_publications.

State v. Gordon, 919 N.W.2d 635 (Iowa Ct. App. 2018).

State v. Guise, 919 N.W.2d 635 (Iowa Ct. App. 2018).

State v. Loomis, 881 N.W.2d 749 (Wis. 2016).

Tutt, Andrew. "An FDA for Algorithms." *SSRN Electronic Journal* 69, no. 1 (2016), 83-123. doi:10.2139/ssrn.2747994.

United States v. Carroll Towing Co., 159 F. 2d 169 (2d. Cir. 1947).

University of Pittsburgh. "Guide to Identifying Personally Identifiable Information (PII)." Information Technology. Last modified February 16, 2021. https://www.technology.pitt.edu/help-desk/how-to-documents/guide-identifying-personally-identifiable-information-pii#:~:text=Personal%20identification%20numbers%3A%20social%20security,Personal%20telephone%20numbers.

WA H.B. 1655

WA S.B. 5527