ANTI-DISCRIMINATION LAW'S CYBERNETIC BLACK HOLE

Marc Canellas[*]

Every system is perfectly designed to get the results it gets. Whatever the stated intent, our American systems addressing crime, housing, family regulation, welfare, employment, and others overwhelmingly result in discrimination. Into these discriminatory systems of humans and organizations, we have increasingly incorporated machines, creating "cybernetic" systems. These cybernetic systems have helped achieve the peak of our devotion to a colorblind society at the expense of a substantively equitable society. Machines increase the speed, scale, and efficiency of operations, while their complex inner-workings and our interdependence on them effectively shield our consciousness and our laws from the reality that their failures disproportionately affect protected groups.

This Essay reveals how cybernetic systems have created black holes in antidiscrimination law where no plaintiff can prove liability no matter the scope of the harm. The fundamental assumptions within antidiscrimination law for disparate treatment and disparate impact cannot capture cybernetic system discrimination. The interdependence of cybernetic systems has prevented experts, let alone plaintiffs, from being able accurately label the human or the machine alone as the sole contributor to failure while complexity means that preventing or identifying a failure requires extensive investigations to determine even general contributors, assuming the point of failure can ever be identified.

When searching for solutions to this black hole problem, it is clear that neither tweaks to antidiscrimination law nor pinches of technological magic will remove this black hole. The way forward is to redefine intentional discrimination as occurring when someone (i) intentionally deployed a system that (ii) caused discrimination against someone based on their protected class. Strict liability is required because it is impossible for plaintiffs to enforce the obligations necessary to prevent discrimination and enables courts to infer intentional discrimination from the intentional deployment of discriminating cybernetic systems. Given that liability then turns on identifying whether there was sufficient discrimination, the law will have to do what America seems to fear most: honestly face the disparities between races, genders, sexual orientation, and others in our society and determine what we are willing to tolerate – an antisubordination lens.

[*] Assistant Public Defender for Office of the Public Defender for Arlington County & the City of Falls Church, Virginia. Chair of the IEEE-USA AI Policy Committee. J.D. from the New York University School of Law and Ph.D. in aerospace and cognitive engineering from the Georgia Institute of Technology. Thank you to Randy Milch, Helen Scott, Jerry Rosenfeld, and the members NYU Cyber Scholars Workshop and the NYU Business Projects Seminar for their helpful comments on earlier versions of this paper. This version is for WeRobot 2021 in September 2021 in Coral Gables, FL, hosted by the University of Miami School of Law.

TABLE OF CONTENTS

I. Introduction

"Every system is perfectly designed to get the results it gets."[1] Let us see the results of our American democratic system: "African Americans are more likely than white Americans to be arrested; once arrested, they are more likely to be convicted; and once convicted, and they are more likely to experience lengthy prison sentences."[2] People of color disproportionately experience homelessness and neighborhoods with more renters of color face higher rates of eviction.[3] African-American children and American Indian and Alaska Native children account for 14% and 1% of the children's population but 23% and 2% of the children in foster care.[4] Typical white families have eight times the wealth of a typical black family and five times the wealth of a typical Hispanic family.[5] White people with more than 16 years of education lived 14.2 years longer than black people with less than 12 years of education.[6] Women consistently earn less than men and the gap is wider for most women of color.[7]

Our American systems addressing crime, housing, family regulation, welfare, employment, and others have increasingly incorporated machines made of mathematics, statistics, and, sometimes, science, into their mix of humans and bureaucracy. We are constructing "cybernetic" systems where humans and machines work together to make decisions within an organization. But is not a revolution but an evolution.[8] Our use of machines represents the peak of evolution, the perfect design, for our devotion to a colorblind society at the expense of a substantively equitable society: do whatever you want, just as long as you do not talk about race, color, gender, religion, or national origin. Machines increase the speed, scale, and efficiency of operations, while their

[1] Susan Carr, *A Quotation with a Life of Its Own*, PATIENT SAFETY & QUALITY HEALTHCARE (July 1, 2008), https://www.psqh.com/analysis/editor-s-notebook-a-quotation-with-a-life-of-its-own/ (attributing the quote to Dr. Paul Batalden).

[2] SENTENCING PROJECT, REPORT OF THE SENTENCING PROJECT TO THE UNITED NATIONS SPECIAL RAPPORTEUR ON CONTEMPORARY FORMS OF RACISM, RACIAL DISCRIMINATION, XENOPHOBIA, AND RELATED INTOLERANCE REGARDING RACIAL DISPARITIES IN THE UNITED STATES CRIMINAL JUSTICE SYSTEM 1 (Mar. 2018) https://www.sentencingproject.org/publications/un-report-on-racial-disparities/

[3] Jaboa Lake, *The Pandemic Has Exacerbated Housing Instability for Renters of Color*, CENTER FOR AMERICAN PROGRESS (Oct. 30, 2020), https://www.americanprogress.org/issues/poverty/reports/2020/10/30/492606/pandemic-exacerbated-housing-instability-renters-color/

[4] *Disproportionality and Race Equity in Child Welfare*, NATIONAL CONFERENCE OF STATE LEGISLATURES (Jan. 26, 2021) (accessed May 27, 2021) https://www.ncsl.org/research/human-services/disproportionality-and-race-equity-in-child-welfare.aspx

[5] Neil Bhutta, Andrew C. Chang, Lisa J. Dettling, and Joanne W. Hsu, *Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances*, FEDS Notes, BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM (2020) https://www.federalreserve.gov/econres/notes/feds-notes/disparities-in-wealth-by-race-and-ethnicity-in-the-2019-survey-of-consumer-finances-20200928.htm

[6] Claire Conway, *Poor Health: When Poverty Becomes Disease*, UCSF MAGAZINE (Fall 2015) https://www.ucsf.edu/news/2016/01/401251/poor-health-when-poverty-becomes-disease

[7] Robin Bleiweis, *Quick Facts About the Gender Wage Gap* (Mar. 24, 2020) https://www.americanprogress.org/issues/women/reports/2020/03/24/482141/quick-facts-gender-wage-gap/

[8] VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 37 (2018). ("The advocates of automated and algorithmic approaches to public services often describe the new generation of digital tools as 'disruptive.' They tell us that big data shakes up hidebound bureaucracies, stimulates innovative solutions, and increases transparency. But when we focus on the programs specifically targeted at the poor and working-class people, the new regime of data analytics is more evolution. It is simply an expansion and continuation of the moralistic and punitive poverty management strategies that have been with us since the 1820's.")

complex inner-workings and our complex interactions with them effectively shields our consciousness and our laws from the reality that their failures disproportionately affect protected groups. We have used machines to "reframe shared social decisions about who we are and who we want to be [into] system engineering problems."[9] And once our social problems are reframed into engineering problems, they are no longer cognizable as discrimination by the laws of a colorblind society. These cybernetic systems are constantly failing to avoid discriminatory outcomes, but they fail in ways that colorblindness cannot even conceive of, leaving people without protection. In our modern cybernetic world, only through casting aside our devotion to colorblindness, talking about and explicitly considering the disproportionate effects on people of different race, color, gender, religion, or national origin can we achieve the equitable society we purportedly aspire to.

This Essay will show that this choice to use colorblindness as the goal of antidiscrimination makes proving discrimination in our modern cybernetic world nearly impossible. However, it is critical to understand that this colorblindness is a choice – a choice of how to understand discrimination and the operative meaning of the Equal Protection Clause of the Fourteenth Amendment to the United States Constitution[10] and its related antidiscrimination statutes addressing employment, of age, disabilities, and housing.[11] The language of "colorblindness" is a shorthand for the anticlassification principle, "holding that the responsibility of the law is to eliminate the unfairness individuals in certain protected classes experience due to decision makers' choices."[12] Answering the question of how to stop discrimination on the basis of race, Chief Justice John Roberts, Jr., replied with the famous anticlassification perspective that dominates American antidiscrimination jurisprudence, "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race."[13] In other words, discrimination on the basis of a protected status will end as soon as the government or other decision makers stop classifying people based on their protected status.[14]

Our society, or at least our judiciary and political leaders, could just as easily have chosen the antisubordination principle which "holds that the goal of antidiscrimination law is, or at least should be, to eliminate status-based inequality due to membership in those classes, not as a matter of procedure, but of substance."[15] At later reply from Justice Sonia Sotomayor countered Chief Justice Roberts' simplistic analysis with an antisubordination perspective: "The way to stop discrimination on the basis of race is to speak openly and candidly on the subject of race, and to apply the Constitution with eyes wide open to the unfortunate effects of centuries of racial

[9] *Id.* at 12.

[10] See U.S. CONST. amend. XIV, § I ("No State shall ... deny to any person within its jurisdiction the equal protection of the laws.").

[11] Title VII of the Civil Rights Act of 1964 (Title VII), 42 U.S.C. § 2000e *et seq.*; the Americans with Disabilities Act of 1990 (ADA), 42 U.S.C. §§ 12112 (b)(2), (b)(6); the Age Discrimination in Employment Act (ADEA) 29 U.S.C. §621 *et seq.*; and, the Fair Housing Act (FHA), 42 U.S.C. §§ 804(a) and 805(a).

[12] Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. LAW REV. 671–732, 723 (2016). (citing Helen Norton, *The Supreme Court's Post-Racial Turn Towards a Zero-Sum Understanding of Equality*, 52 WILLIAM MARY LAW REV. 197, 206, 209 (2010).)

[13] *Parents Involved in Community Schools v. Seattle School District No. 1*, 551 U.S. 701, 748 (2007) (plurality opinion).

[14] Ronald Turner, *The Way to Stop Discrimination on the Basis of Race...*, 11 STANF. J. CIV. RIGHTS CIV. LIB. 45, 47 (2015).

[15] Barocas and Selbst, *supra* note 13 at 723. (citing Norton, *supra* note 13 at 206, 209.)

discrimination."[16] In other words, antidiscrimination law and the courts adjudicating it "should engage in an analysis that is cognizant of the actuality and effects of [discrimination] as evidenced by the lived experiences of those historically subjected to and affected by the legal and social practice of racism-based subordination."[17] So while anti-classification only regulates processes, anti-subordination instead focuses on the outcomes. As this Essay will show, the inner processes of modern discrimination are nearly impossible to govern, leaving anti-subordination as the only option to truly address discrimination.

Legal scholars, engineers, and scientists have long shown that anti-discrimination law premised on anticlassification is inadequate to address the most virulent forms of discrimination. This Essay leverages the scientific and engineering understanding of cybernetics to integrate and build upon the three major scientifically based challenges levied against anticlassification which are based on psychological science, social institutions, and machines and algorithms. First, scholars have long used psychological science to show that the law's focus on intentional discrimination simply ignores the reality that discrimination is often the product of something less than "discriminatory animus"[18] – referred to in this Essay as "human" discrimination. The ideals

[16] *Schuette v. Coalition to Defend Affirmative Action, Integration and Immigration Rights and Fight for Equality by Any Means Necessary*, 572 U.S. 291, 381 (2014) (Sotomayor, J., dissenting) ("In my colleagues' view, examining the racial impact of legislation only perpetuates racial discrimination. This refusal to accept the stark reality that race matters is regrettable. The way to stop discrimination on the basis of race is to speak openly and candidly on the subject of race, and to apply the Constitution with eyes open to the unfortunate effects of centuries of racial discrimination. As members of the judiciary tasked with intervening to carry out the guarantee of equal protection, we ought not sit back and wish away, rather than confront, the racial inequality that exists in our society. It is this view that works harm, by perpetuating the facile notion that what makes race matter is acknowledging the simple truth that race does matter.")

[17] Turner, *supra* note 15 at 47.

[18] *Price Waterhouse v. Hopkins*, 490 U.S. 228, 276 (Justice O'Connor explaining that the key inference is about if the "employer's discriminatory animus made a difference to the outcome"); Charles R. Lawrence, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, 39 STANFORD LAW REV. 317, 321–22 (1987). (opposing the language of "unintentional" discrimination: given that "the illness of racism infects almost everyone. … I argue that this is a false dichotomy. Traditional notions of intent do not reflect the fact that decisions about racial matters are influenced in large part by factors that can be characterized as neither intentional-in the sense that certain outcomes are self-consciously sought nor unintentional-in the sense that the outcomes are random, fortuitous, and uninfluenced by the decisionmaker's beliefs, desires, and wishes."); John Tyler Clemons, *Blind Injustice: The Supreme Court, Implicit Racial Bias, and the Racial Disparity in the Criminal Justice System*, 51 AM. CRIM. LAW REV. 689, 689–90 (2014). ("Decades of psychological research has demonstrated that the most insidious form of racial bias is actually implicit and subconscious… Moreover, research has consistently shown that such racial bias--termed 'implicit racial bias' by the psychological literature--is capable of affecting conscious behavior and exists independently of individuals' conscious and explicit beliefs about racial equality. By clinging to an outdated and incomplete definition of racial discrimination, the Court has made a series of decisions that have permitted and exacerbated the damage that implicit racial bias wreaks on racial minorities. …[T]he Court has rejected one of its most powerful tools for controlling the effects of such bias, spurning disparate impact theory in favor of an intent-based standard that is all but impossible for plaintiffs to meet.") Mona Lynch & Craig Haney, *Looking Across the Empathic Divide: Racialized Decision Making on the Capital Jury*, 2011 MICH. STATE LAW REV. 573, 606–07 (2011). ("remedies [for discrimination] must be premised on what we now know about the nature and operation of modern racism, including the frank recognition that, contrary to prevailing legal wisdom, it is not solely a problem of conscious, motivated individual actors who engage in 'purposeful discrimination.' Because it often operates implicitly, as a function of structural, institutional, and even biographical forces, it must be combated with remedies that extend far beyond the openly prejudiced, single individuals who are most often targeted.") Amelia M. Wirts, *Discriminatory Intent and Implicit Bias: Title VII Liability for Unwitting Discrimination*, 58 BOSTON COLL. LAW REV. 809 (2017); Julia Kobick, *Discriminatory Intent Reconsidered: Folk Concepts of Intentionality and Equal Protection Jurisprudence*, 45 HARV. CIV. RIGHTS-CIV. LIB. LAW REV. 517 (2010).

embodied in anticlassification attempting to eliminate intentional and overt discrimination have only caused us to reject that specific strain of discrimination leaving equally powerful sources of discrimination and their discriminatory outcomes untouched. Our subconscious, unconscious, or implicit biases, and microagressions[19] are still ingrained in us by our society, experiences, and relationships.[20] A child need not need to be told that a Black person, a woman, a person with low income, or a person with mental disabilities is inferior, they learn that lesson clearly by observing how others treat them. These prejudices and biases affect decision making but because they are tacit and unarticulated, they are not experienced at the conscious level.[21] However, it is not only that people naturally absorb and act upon discriminatory motives without their conscious mind, but that the cognitive processes of grouping and differentiating, also known as stereotyping, are normal cognitive processes.[22] Stereotyping is a cognitive process, not motivational, such that stereotypes can "operate absent intent to favor or disfavor members of a particular social group."[23] None of these less-than-fully-conscious biases are cognizable by antidiscrimination law using anticlassification principles.

As evidenced by the description above, individual discrimination is scientifically understood to be, in part, a product of the discrimination embedded in a society's policies. This set of policies will be referred to as systemic discrimination, grouping together institutional discrimination (intentionally discriminatory policies) and structural discrimination (unintentionally discriminatory policies).[24] These terms are often used interchangeably as shorthand for "any sort of discrimination produced by large-scale, stable social arrangements, whether generated through individual action (bigoted or otherwise) or not."[25] However, just as psychology has rejected the legal formalism that individual discrimination is either intentional or unintentional, so too have social scientists rejected the theories that systemic discrimination is either the sole product or independent of individually-motivated discrimination.[26] Following the

[19] Peggy C. Davis, *Law as Microaggression*, 98 YALE LAW J. 1559, 1576 (1989). (defining microagressions as "stunning, automatic acts of disregard that stem from unconscious attitudes of white superiority and constitute a verification of black inferiority." *Id.* at 1572 n. 59. (Microagressions are key "mechanisms of contemporary racialism" built on "cognitive habit, history, and culture.")

[20] Lawrence, *supra* note 19 at 323. ("In short, requiring proof of conscious or intentional motivation as a prerequisite to constitutional recognition that a decision is race-dependent ignores much of what we understand about how the human mind works. It also disregards both the irrationality of racism and the profound effect that the history of American race relations has had on the individual and collective unconscious.") See also, Sheri Lynn Johnson, *Unconscious Racism and the Criminal Law*, 73 CORNELL LAW REV. 1016, 1019 (1988). ("The concept of purposeful discrimination, or at least its terminology, does not mesh well with unconscious race discrimination.") Richard A. Wasserstrom, *Racism, Sexism, and Preferential Treatment. An Approach to the Topics*, 24 UCLA LAW REV. 581, 590 (1977).

[21] Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STANFORD LAW REV. 1161, 1165 (1995). ("I conclude that, while the assumptions undergirding disparate treatment theory generally reflect the thinking about intergroup bias and human inference accepted into the 1970s, these assumptions have been so undermined, both empirically and theoretically, that they can no longer be considered valid.")

[22] *Id.* at 1187. (discussing social cognition theory)

[23] *Id.* at 1188.

[24] Fred L. Pincus, *Discrimination Comes in Many Forms: Individual, Institutional, and Structural*, 40 AM. BEHAV. SCI. 186, 186 (1996).

[25] Ian F. Haney Lopez, *Institutional Racism: Judicial Conduct and a New Theory of Racial Discrimination*, 109 YALE LAW J. 1717, 1727 n. 32 (2000).

[26] *Id.* at 1727. (using the theory of "new institutions" to define "institutional racism" as a third way to understand systemic discrimination that "neither relies on motivated behavior nor dismisses behavior altogether, but rather one

psychological perspective that “nonconscious [discriminatory] beliefs permeate society,”[27] systemic discrimination explains that discrimination can be the product of group, social and organizational prejudices that coalesce into systems and institutions,[28] embed into organizational and workplace dynamics,[29] and then advance without conscious intent and without singular discrete and consequential decisions.[30] Ian Haney López links individual psychological discrimination with institutional and systemic discrimination by “restat[ing] in institutional language Charles Lawrence’s [psychological] observation that ‘we are all racists’: In this country we are all constituted by and cognitively rely on racial institutions.”[31] Here again, scholars agree that antidiscrimination law premised on anticlassification is wholly ill-equipped to address institutional or systemic discrimination.[32]

To this constellation of colorblind antidiscrimination law’s failures, researchers have recently added a paradigm that ultimately surfaces near-identical issues: machine discrimination. Here, machines are the instruments, typically based in mathematics, statistics, or science, that inform or control decisions that affect people’s lives.[33] They can be hiring algorithms,[34] child welfare algorithms,[35] face recognition systems,[36] or probabilistic DNA software.[37] But it is critical

that focuses on the sort of nonintentional behavior emphasized by institutional analysis”)

[27] *Id.* at 1808.

[28] *Id.* at 1808.

[29] Tristin K Green, *Discrimination in Workplace Dynamics: Toward a Structural Account of Disparate Treatment Theory*, 38 HARV. CIV. RIGHTS-CIV. LIB. LAW REV. 91, 92 (2003). *See,* Martha Chamallas, *Structuralist and Cultural Domination Theories Meet Title VII: Some Contemporary Influences*, 92 MICH. LAW REV. 2370 (1994); Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUMBIA LAW REV. 458 (2001).

[30] Samuel R Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CALIF. LAW REV. 1, 12–14 (2006).

[31] Lopez, *supra* note 26 at 1808–09.

[32] *Id.* at 1730. (“institutional analysis demonstrates that the current Supreme Court's reasoning is exactly backward: Racism occurs frequently–and perhaps predominantly–without any specific invocation of race, while the explicit consideration of race may have as its aim racism's amelioration rather than perpetuation.”); Bagenstos, *supra* note 31 at 3. (“These difficulties are mere symptoms of a deeper problem: structural employment inequalities cannot be solved without going beyond the generally accepted normative underpinnings of antidiscrimination law. Because courts and legislatures have proven unable or unwilling to take that step, structural discrimination advocates essentially proceed by indirection. They seek to develop rules that will empower workplace constituencies who will internalize and advance the correct vision of equality. But unless courts have some normative idea of what workplace equality should mean, they will be unable to ensure that those workplace constituencies will serve the purposes of antidiscrimination law.”) Randall L. Kennedy, *McClesky v. Kemp: Race, Capital Punishment, and the Supreme Court*, 101 HARV. LAW REV. 1388, 1442–43 (1988). (describing systemic racism as another form of discrimination that the Supreme Court typically deems non-justiciable).

[33] This is a broader perspective of machine than most scholars who are often narrowly focused on big data, machine learning or artificial intelligence. Nevertheless, for summaries of the components of machine discrimination *see*, Barocas and Selbst, *supra* note 13.; David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 UNIV. CALIF. DAVIS LAW REV. 653, 671 (2017).

[34] Jeffrey Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*, REUTERS (Oct. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[35] Dan Hurley, *Can an Algorithm tell when Kids are in Danger?*, N.Y. TIMES MAGAZINE (Jan 2, 2008) https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html

[36] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 *in* PROCEDDINGS OF MACHINE LEARNING RESEARCH 1 (2018).

[37] Lauren Kirchner, *Traces of Crime: How New York’s DNA Techniques Became Tainted*, NEW YORK TIMES, September 4, 2017, https://www.nytimes.com/2017/09/04/nyregion/dna-analysis-evidence-new-york-disputed-techniques.html; Marc Canellas, *Defending IEEE Standards in Federal Criminal Court*, 54 IEEE COMPUT. (2021).

to realize that the major concerns of machine discrimination are not new and do not depend on data being "big," the machine being "intelligent," or the machine truly being a "machine" at all. Nothing more than basic principles of adding and dividing were needed to leverage the 1840 U.S. Census data to calculate that northern free Black people were nearly ten times more likely to be classified as insane than southern Black people and then use that calculation to justify the belief that slavery had "a wonderful influence upon the development of the moral faculties and the intellectual powers" of Black people.[38]

The conclusions of those studying machine discrimination align with those studying human and systemic discrimination: antidiscrimination law is largely incompatible with the reality of machine discrimination. Machine discrimination is understood best as something neither intentional nor unintentional. Whether the humans designing and deploying these machines are clueless or careless, the methods of machine learning "can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society."[39] In addition, the speed, scale, and sophistication of machines can "systematize and conceal discrimination."[40] Rather than affecting one employment decision or one criminal defendant, a single machine can affect thousands, all while hiding its true mechanisms in a code that few understand. Unsurprisingly, given that machines are a product of and operate in a world of human psychological and systemic discrimination, machine discrimination is also understood to be largely beyond the capabilities or willingness of current courts and antidiscrimination law.[41]

This Essay introduces the theory of cybernetic system discrimination, integrating these three paradigms of human, systemic, and machine discrimination into a single framework. The term *cybernetic* refers to the belief that "society can only be understood through a study of… messages between man and machines, between machines and man, and between machine and machine."[42] In our cybernetic world where humans and machines make decisions together, the "messages between" are the critical source for understanding why they fail, or, in other words, why they discriminate. Interaction defines us. Machines are the medium through which we humans interact with the world and thus the medium through which our society operates – from hiring and

[38] Edward Jarvis, *Statistics of Insanity in the United States*, 27 BOSTON MED. SURG. J. 116, 119 (1842). ("Slavery has a wonderful influence upon the development of moral faculties and the intellectual powers; and refusing man many of the hopes and responsibilities which the free, self-thinking and self-acting enjoy and sustain, of course it saves him from some of the liabilities and dangers of active self-direction"). This story is summarized by Ibram Kendi who further explains, unsurprisingly, that Dr. Jarvis, who was an antislavery activist, later "found errors everywhere [in the 1840 Census]. Some northern towns reported more Black lunatics than Black residents." IBRAM X. KENDI, STAMPED FROM THE BEGINNING: THE DEFINITIVE HISTORY OF RACIST IDEAS IN AMERICA 180–81 (2017).

[39] Barocas and Selbst, *supra* note 13 at 671. See generally, Anupam Chander, *The Racist Algorithm?*, 115 MICH. LAW REV. 1023 (2017); Pauline T Kim, *Data-Driven Discrimination at Work*, 58 WILLIAM MARY LAW REV. 857 (2017); Margaret Hu, *Algorithmic Jim Crow*, 86 FORDHAM LAW REV. 633 (2017); Pauline T Kim, *Auditing Algorithms for Discrimination*, 166 UNIV. PA. LAW REV. ONLINE 189 (2017); Danielle Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. LAW REV. 1–33 (2014).

[40] Joshua A Kroll et al., *Accountable Algorithms*, 165 UNIV. PA. LAW REV. 633, 680 (2017).

[41] Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J. LAW TECHNOL. 889 (2018); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE LAW J. 1043 (2019); Deven R Desai & Joshua A Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J. LAW TECHNOL. 1 (2017); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. LAW REV. 109 (2017); Jason R Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEORGETOWN LAW J. 803 (2020).

[42] NORBERT WEINER, THE HUMAN USE OF HUMAN BEINGS: CYBERNETICS AND SOCIETY 16 (1988).

firing to administering justice and disbursing welfare. *System* refers to the recognition that humans are making these decisions within organizational structures (the social system), utilizing machines (the technical system) to achieve overall system goals and objectives.[43] "These systems involve context-rich workplace settings, organizational structure, human operators, and sophisticated technology that when taken collectively are known as complex sociotechnical systems."[44] As a result, accidents or discrimination can result "from dysfunctional interactions among system components," not just component failures.[45] In sum, the causes of accidents or discrimination are often a product of the interactions between the individual psychology of the human (human discrimination), the machine (machine discrimination), and the broader organization (systemic discrimination). Focusing too much on the human, the machine, or the organization ignores these interactions, leaving the interdependent complexities of failure unexplored.[46]

From this framework, the theory of cybernetic system discrimination defines discrimination as actions that, due to the absence or failure of barriers and controls, produce injuries to persons because of their race, color, gender, religion, or national origin. This definition is analogized from the science of system safety and accident causation which specializes in investigating and understanding why catastrophic failures occur: from why airplanes crash to why nuclear plants meltdown.[47] A discrimination lawsuit is in effect a demand for an investigation into a different type of catastrophic failure: a system's failure to not discriminate when hiring, prosecuting, disbursing benefits, among many others.

The failure of antidiscrimination law and the colorblind anticlassification perspective is caused by the belief that discrimination can only occur due to a specifically identifiable failure of human self-control or a machine design. Antidiscrimination law simply cannot cope with the

[43] Marc C. Canellas et al., *Framing Human-Automation Regulation: A New Modus Operandi from Cognitive Engineering*, *in* WEROBOT 2017, 6 (2017).

[44] Marc C. Canellas et al., *Framing Human-Automation Regulation: A New Modus Operandi from Cognitive Engineering*, *in* WEROBOT 2017, 6 (2017). For complex sociotechnical systems, *see generally*, Gordon Baxter & Ian Sommerville, *Socio-technical systems: From design methods to systems engineering*, 23 INTERACT. COMPUT. 4–17 (2011); Guy H Walker et al., *A review of sociotechnical systems theory: a classic concept for new command and control paradigms*, 9 THEOR. ISSUES ERGON. SCI. 479–499 (2008); Patrick Waterson et al., *Defining the methodological challenges and opportunities for an effective science of sociotechnical systems and safety*, 58 ERGONOMICS 565–599 (2015). For cognitive systems engineering, *see generally*, Erik Hollnagel & David D Woods, *Cognitive Systems Engineering: New Wine in New Bottles*, 18 INT. J. MAN-MACH. STUD. 583–600 (1983); JENS RASMUSSEN, ANNELISE MARK PEJTERSEN & L P GOODSTEIN, COGNITIVE SYSTEMS ENGINEERING (1994); DAVID D. WOODS & ERIK HOLLNAGEL, JOINT COGNITIVE SYSTEMS: PATTERNS IN COGNITIVE SYSTEMS ENGINEERING (2006); David D Woods & Emilie M Roth, *Cognitive engineering: Human problem solving with tools*, 30 HUM. FACTORS J. HUM. FACTORS ERGON. SOC. 415–430 (1988).

[45] Joseph H. Saleh & Cynthia C. Pendley, *From learning from accidents to teaching about accident causation and prevention: Multidisciplinary education and safety literacy for all engineering students*, 99 RELIAB. ENG. SYST. SAF. 105–113, 105 (2012). (citation omitted)

[46] Marc Canellas & Rachel Haga, *Unsafe at Any Level*, 63 COMMUN ACM 31–34 (2020); Matthew J. Miller & Karen M. Feigh, *Addressing the envisioned world problem: a case study in human spaceflight operations*, 5 DES. SCI. 1 (2019); Karen M Feigh & Amy R Pritchett, *Requirements for Effective Function Allocation A Critical Review*, 8 J. COGN. ENG. DECIS. MAK. 23–32 (2014); Matthew Johnson et al., *Coactive Design: Designing Support for Interdependence in Joint Activity*, 3 J. HUM.-ROBOT INTERACT. 43–69 (2014).

[47] Saleh and Pendley, *supra* note 46 at 105. (quoting the Department of Energy, which defines an accident as an ‘‘unwanted transfer [or release] of energy that, due to the absence or failure of barriers and controls, produces injury to persons, damage to property, or reduction in process output’’)

realities of cybernetic system discrimination that scientists and engineers have understood for decades. James Reason, one of the foremost researchers of system safety and accident causation, explained that as early as the 1990's,

> "neither investigators nor responsible organizations are likely to end their search for causes of an organizational accident with the mere identification of 'sharp-end' human failures. Such unsafe acts are now seen more as consequences than as principle causes. … Although fallibility is an inescapable part of the human condition, it is now recognized that people working in complex systems make errors or violate procedures for reasons that generally go beyond the scope of individual psychology. These reasons are *latent conditions*.
>
> Latent conditions are to technological organizations what resident pathogens are to the human body. Like pathogens, latent conditions – such as poor design, gaps in supervision, undetected manufacturing defects or maintenance failures, unworkable procedures, clumsy automation, shortfalls in training, less than adequate tools and equipment – may be present for many years before they combine with local circumstances and active failures to penetrate the system's many layers of defences. They arise from strategic and other top-level decisions made by governments, regulators, manufacturers, designers and organizational managers. The impact of these decisions spreads throughout the organization, shaping a distinctive corporate culture… and creating error-producing factors within the individual workplaces."[48]

Because antidiscrimination law's current anticlassification perspective is unable to capture how cybernetic systems fail and discriminate, cybernetic systems have become a black hole for the law. Antidiscrimination law is completely incapable of identifying discrimination or assigning liability. The law therefore stands opposed to how lay people and experts understand and experience discrimination, actively exacerbating rather than reducing tension.[49] Many of those discriminated against will have insufficient remedies, told that the discrimination they face is not recognized under the law, that the discrimination they face is not meaningful or real. Even in those few situations where the court does recognize discrimination, the desire to blame discrimination solely on a human with discriminatory animus or a flawed machine will still exacerbate tensions. Cybernetic system discrimination is the product of joint human-machine decision making within a broader organization. The party held accountable will feel as though they are the moral crumple zone, taking on the blame for errors or accidents not entirely in their control,[50] while the other party who avoids accountability will lack any incentive to remedy their contributions to discrimination. This Essay attempts to find a way forward.

Section II of this Essay identifies the six components of the cybernetic system framework integrating human, machine, and systemic (organizational) sources of discrimination. The theoretical process of a cybernetic system describes situations where a human uses a machine to

[48] JAMES REASON, MANAGING THE RISKS OF ORGANIZATIONAL ACCIDENTS 10 (1997).

[49] Krieger, *supra* note 22 at 1238.

[50] Madeleine Clare Elish, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, 5 ENGAG. SCI. TECHNOL. SOC. 40, 42 (2019).

make a decision such as a human resources manager using a hiring software to decide who to hire. First, (i) *inputs* like job applications are fed into the (ii) *machine*, hiring software, which then (iii) *interacts* with the human by presenting information or recommendations. The (iv) *user*, a manager intakes that information and then can (v) provide additional information to the machine or modify the machine through a process known as *feedback* before making the final decision. All these components exist within an (vi) *organization*, the organization providing training, procedures, processes, and pressures influencing each component.

Section III shows how existence and interaction of these components within cybernetic systems result in interdependence and complexity, two primary characteristics of cybernetic systems at odds with current antidiscrimination law and the anticlassification perspective reliant on the myths of deconstruction and dualism. Deconstruction is the belief that if only we had enough evidence and access to the human, machine, and organization, then we could get inside the head or code of the discriminator to determine what really happened; [51] and dualism is the belief that discrimination is either due to human or machine actors, one or the other, nothing in between.[52] Compare these myths to the realities of interdependence where each human or machine action depends on the actions of the other[53] and complexity where there are numerous components determining the ultimate performance or failure of the cybernetic systems in ways that are difficult to measure, predict, and control.[54] Therefore, whenever a cybernetic system discriminates, interdependence means that one can rarely accurately label the human or the machine alone as the sole contributor to failure while complexity means that preventing or identifying a failure requires extensive investigations to determine even general contributors, assuming the point of failure can ever be identified. These two characteristics are the ultimate downfall of anticlassification as a rational premise for antidiscrimination.

Section IV goes further to identify the assumptions implicit in the two types of prohibited discrimination – intentional discrimination (disparate treatment) and unintentional discrimination (disparate impact) – to show how they, too, specifically conflict with the reality of cybernetic system discrimination characterized by interdependence and complexity. The Supreme Court has defined the discriminatory intent and disparate treatment tests to address intentional discrimination, but in cybernetic systems, the amount of discrimination that the Court deems justiciable reduces to a null set.[55] This analysis focuses primarily on employment discrimination and Title VII as an exemplar of antidiscrimination law because of the highly-developed caselaw and legal scholarship, the increasing use of machines in employment decision making, and the recognition of disparate impact which is often argued as a potential solution to the issues with

[51] SIDNEY DEKKER, TEN QUESTIONS ABOUT HUMAN ERROR: A NEW VIEW OF HUMAN FACTORS AND SYSTEM SAFETY 2 (2004).

[52] *Id.* at 2–3.

[53] Matthew Johnson et al., *Coactive Design: Designing Support for Interdependence in Joint Activity*, 3 J. HUM.-ROBOT INTERACT. 43–69, 47 (2014). ("Interdependence" describes the set of complementary relationships that two or more parties rely on to manage required (hard) or opportunistic (soft) dependencies in joint activity.)

[54] K VICENTE, COGNITIVE WORK ANALYSIS : TOWARD SAFE, PRODUCTIVE, AND HEALTHY COMPUTER-BASED WORK 14 (1999).,

[55] See, e.g., Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. LAW REV. 671, 698, 701 (2016). (Within the scholarly literature, there is surprising unanimity that the law does not adequately address unconscious disparate treatment. … In sum, aside from rational racism and masking (with some difficulties), disparate treatment doctrine does not appear to do much to regulate discriminatory data mining.") (internal citation and quotation omitted).

anticlassification.[56] Disparate treatment understands discrimination as occurring when a human, (i) at the moment of the decision, (ii) in complete control of their decision-making process, (iii) intentionally, and (iv) invidiously discriminates against someone based on their protected class. Disparate impact understands discrimination as (i) an unfair machine or human's test has discriminatory results (ii) caused by a specific failure, (iii) identifiable prior to deployment, (iv) and that there is an equally effective, less discriminatory alternative employment practice which the employer refused to adopt. Complexity raises the difficulty of finding sufficient evidence to prove any of these elements while interdependency obfuscates any ability to make sense of what is found. The sum of legal and technical barriers that plaintiffs must overcome to show disparate treatment or disparate impact transforms cybernetic systems into black holes for antidiscrimination analysis – a place to where discrimination can occur without the law ever recognizing it. This reality means that disparate treatment and disparate impact cannot contemplate cybernetic system discrimination leaving those discriminated against without remedy and those even thinking about avoiding discrimination without any incentive to do so. In other words, the anticlassification perspective sanctions discrimination in our cybernetic world, and moreover, incentivizes users and organizations to adopt even more cybernetic systems in order to shield themselves from liability.

Section V searches for solutions. A review of the numerous proposed reforms shows that almost all singularly focus on individual elements of the cybernetic system, without accounting for interdependence and complexity, making them valuable but insufficient to address cybernetic system discrimination. The reality is that cybernetic system discrimination cannot be found by the limited adjudication methods allowed by colorblind anticlassification. The only true methods to address cybernetic system discrimination depend upon antisubordination perspective of antidiscrimination law[57] enforced through strict liability. Specifically, I propose a new understanding of intentional discrimination as occurring when someone (i) intentionally deploys a system that (ii) causes discrimination against someone based on their protected class. Once sufficient identification is identified as the output of a system, then the law should presume intent on behalf of the person or persons deploying the system. This strict liability enforcement should not be surprising as it is standard practice when those harmed cannot adequately enforce the obligations necessary to ensure reasonable quality control.[58] With cybernetic systems, plaintiffs who have been discriminated against did not choose to deploy the complex, interdependent system nor did they have did not control the operation of the system, and they are certainly not as well placed as defendants to determine how to ensure future systems do not discriminate. Moreover, because this puts all the pressure on the question of identifying discriminatory outcomes, it demands the necessary antisubordination conversation that our society desperately needs to have about what amount of "discriminatory misperformance" constitutes discriminatory outcomes. If the expansive literature of human discrimination, systemic discrimination, and machine discrimination were not reason enough to abandon the naïve and oppression-sanctioning

[56] Barocas and Selbst, *supra* note 13; Kim, *supra* note 40; Charles A. Sullivan, *Employing AI*, 63 VILLANOVA LAW REV. 395 (2018); Krieger, *supra* note 22.

[57] Cheryl I Harris & Kimberly West-Faulcon, *Reading Ricci: Whitening Discrimination, Racing Test Fairness*, 58 UCLA LAW REV. 73 (2010). Barocas and Selbst, *supra* note 13 at 726. ("where the internal difficulties cannot be overcome, there is likely no way to correct for the discriminatory outcomes aside from results-focused balancing, and requiring this will pose constitutional problems.")

[58] Mark A. Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation*, 105 CALIF. LAW REV. 1611, 1664 (2017). Strict liability also compelled from a moral perspective to ensure anti-discrimination law is achieves accountability. Wirts, *supra* note 19 at 849–50.

anticlassification principle, one can only hope that cybernetic system discrimination which integrates and lays out the true black-hole nature of these challenges is the final nail in the coffin for anticlassification.

Section VI concludes with the broader implications of the cybernetic system framework: significant parts of our legal system are incompatible with the reality of cybernetic systems and antidiscrimination laws are only one of many areas in need of reexamination. The principles of interdependence and complexity are upending the law and erecting barriers to justice whether we chose to choose to be blind to them or not. Anywhere the law relies on discretion, reasonableness, negligence, or mens rea, cybernetic black holes will be found.

II. How Cybernetic Systems Fail

Cybernetics is founded on the idea that "society can only be understood through a study of… messages between man and machines, between machines and man, and between machine and machine."[59] Though the buzzword technologies – algorithms, artificial intelligence, and robotics – are prevalent, they are simply the latest generation of cyber beings we have used to build our cybernetic world. In this cybernetic world, the "messages between" are the critical components. Interaction defines us. Not only are these cyber beings ever-present in our lives, they are the medium through which we humans interact with the world and thus the medium through which our laws operate.

When humans use machines, they cease being two separate entities acting as the sum of their parts. Instead, they become a new entity, a "cybernetic" system, with unique characteristics all its own. While this theory of cybernetics may seem philosophical, and does, in fact, have many philosophical developments, it is very much a technical framework for how cybernetic systems are designed, built, and deployed today – from computers, mobile phones, websites, to cars, airplanes, and robotics. Numerous disciplines have grown up out of this cybernetic perspective and are used to analyze and design these "messages between": from human factors and cognitive engineering to myriad specialties of human-(machine, robot, computer, automation) interaction. This entity, the cybernetic system, demanded the development of these new disciplines because of its unique properties including interdependence and complexity. Interdependence describes relationships where what one agent does depends on what each other agent does; requiring coordination in time and space, and some amount of transparency and trust. Complexity describes systems where problems are ill-structured, with numerous shifting or ill-defined goals, to be addressed by multiple agents in high stakes environments that are changing over time. These properties create significant barriers for professional researchers and engineers, and as will be shown in this Essay, our administration of law.

This Section first outlines the cybernetic system framework, an element-by-element decomposition of a typical cybernetic system to show how it is constructed of the input, machine, human-machine interaction, user, feedback, and the organization. An example of using Google Maps to get to a new location is used to show how almost all our modern decisions are made as part of cybernetic systems and how the question of accountability for failure can be surprisingly

[59] WEINER, *supra* note 43 at 16.

elusive. Second, this framework is applied to one of the leading employment discrimination cases, *Ricci v. DeStefano*,[60] to show how something even something as seemingly simple as a paper test still fits within the cybernetic framework. Moreover, by reorganizing Justice Ginsburg's dissent, the cybernetic system framework is shown to be a valuable method for plaintiffs or challengers to know where to look within a cybernetic system for potential sources of discrimination and organize their arguments.

A. Cybernetic System Framework

There are six elements of cybernetic systems combine to produce outputs, as shown in Figure 1: the input, machine, human-machine interaction, user, feedback, and the organization, each of which can contribute to failures and discriminatory outcomes alone or jointly. The existence and interaction of these six elements gives cybernetic systems two of their key properties that have long caused trouble for scientists and engineers, for our legal system, too: interdependence (each human or machine action depends on the actions of the other[61]) and complexity (there are numerous components determining the ultimate performance or failure of the cybernetic systems in ways that are difficult to measure, predict, and control[62]). The basic theoretical process of a cybernetic system describes situations where a human uses a machine to make a decision such as a human resources manager using a hiring software to decide who to hire. First, (i) *inputs* like job applications are fed into the (ii) *machine*, hiring software, which then (iii) *interacts* with the human by presenting information or recommendations. The (iv) *user*, a manager intakes that information and then can (v) provide additional information to the machine or modify the machine through a process known as *feedback* before making the final decision. All these components exist within a (vi) *sociotechnical system*, the organization providing training, procedures, processes, and pressures influencing each component.

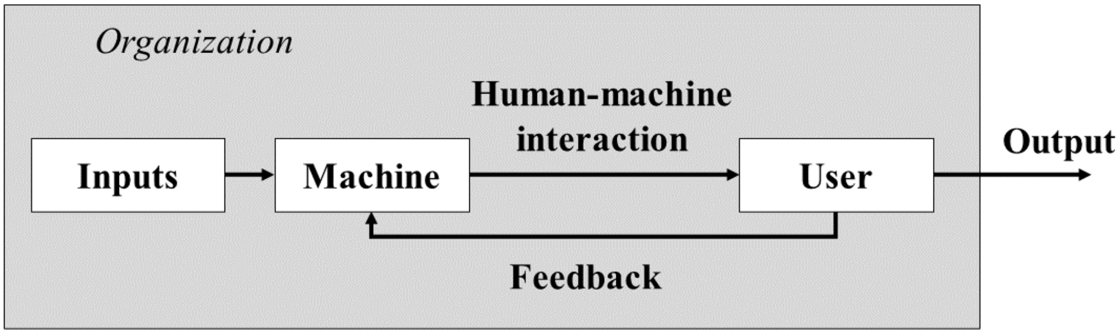

Figure 1. Cybernetic system framework.

As stated above, you are part of multiple cybernetic systems every day. For example, anytime

[60] *Ricci v. DeStefano*, 557 U.S. 557 (2009).

[61] Johnson et al., *supra* note 54 at 47. ("Interdependence" describes the set of complementary relationships that two or more parties rely on to manage required (hard) or opportunistic (soft) dependencies in joint activity.)

[62] VICENTE, *supra* note 55 at 14.,

you use a phone or computer, you become part of a cybernetic system. As shown in Figure 2, recall the times you have used a digital navigation app like Google Maps to travel to a new location. Your goal is to arrive at your destination – that is the *output* you want from your interaction with Google Maps – and you intuitively will use a myriad of complex, interdependent actions to leverage your app (your machine) to get to the destination. First, you *input* your starting location and destination into the *machine*, Google Maps, which uses software to calculate the best route and then *interacts* with you by presenting information or recommendations about the best route. You, the *user*, intake that information to inform your driving and then provide additional information to Google Maps based on your location (i.e., having your GPS "on") through a process known as *feedback* before arriving at your destination, *output*. You are operating Google Maps and navigating within the broader *organization* of rules and environments of buses, cars, subways, pedestrians, and weather which influences every other element, e.g., by controlling when you can provide feedback (losing GPS or internet in tunnels) or what options are even available and how they change (due to delays, changes in traffic, or accidents).

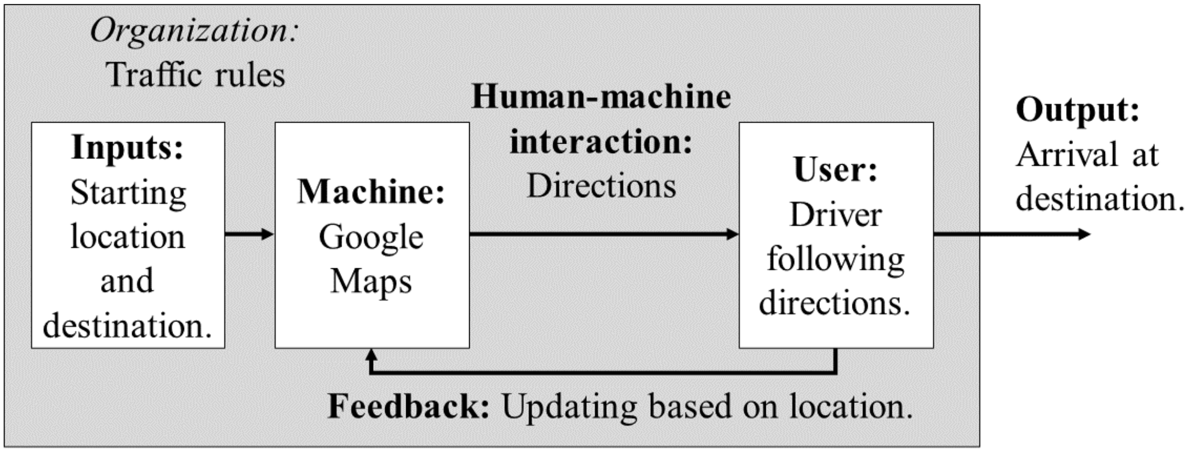
Organization:
Traffic rules
Inputs:
Starting location and destination.
Machine:
Google Maps
Human-machine interaction:
Directions
User:
Driver following directions.
Output:
Arrival at destination.
Feedback: Updating based on location.

Figure 2. Cybernetic framework as applied to using Google Maps for navigation.

You may not notice the nature of the cybernetic systems when everything goes smoothly, but when the system fails to produce the desired output, you will likely immediately be able to identify the various elements, the interdependent complexities. Sometimes, you have followed the directions, only to realize half-way to your destination that the route went completely out-of-the-way or is much more complicated or unreliable than you would have preferred. Sure, you input the correct starting location and destination and told the app to take the "fastest route" but you did not want to go miles out of your way to save a few minutes. Or you realize that the route did not seem to account for the likelihood of traffic or the likelihood that the bus or train would be late (or maybe not arrive at all). When you finally arrived at your destination you offered up the cause of your delay: but for using the app, you would have arrived on time. In the app's defense, if you had not used the app at all, you would never have arrived because you did not know how to get to your destination. And the app's directions may have not been perfect, but they were a viable path towards your destination, not to mention that they were just a recommendation that you willingly requested, accepted and followed. Still, you contend, that the app was not giving you the directions you desired, was not clear about the uncertainty in its time estimates, nor was it accounting for important contexts that were necessary for your decision. Ultimately the determination of responsibility remains elusive: was it solely the app's fault that you arrived late or was it your choice to use the app's recommendation that caused the delay? The responsibility is likely suspended somewhere in between.

These frustrations and complications that you feel are commonly studied in the science and engineering fields related to cybernetics. Maybe the app did not have the proper information necessary to guide you as specifically as you wanted because its inputs were poor quality. Or maybe the inputs were fine but there was something wrong with the models and algorithms inside the app. In the human-machine interaction element, maybe you were "cognitively railroaded" such that you could assess whether the app was operating appropriately due to improper knowledge, information, and time.[63] The app's interface was not sufficiently explainable, transparent or interpretable. Alternatively, in the user element, maybe you were not trained properly in how the app was working, so you overtrusted the app substituting your judgment for the apps. At an organizational level, maybe the allocation of work between you and the app in terms of selecting routes was improper.[64] This is all to say that state-of-the-art research is still grappling with how to predict and govern the outcomes of cybernetic systems ex-ante and assigning causation or responsibility for the outcome ex-post. The law, however, continues moving forward without fully appreciating the consequences of these complexities.

B. Applying the Cybernetic System Framework

The best way to understand how this cybernetic framework can be used in legal proceedings to make arguments that cybernetic systems are in fact inadequate and likely to produce discriminatory outcomes is to look at one of the leading cases in employment discrimination, *Ricci v. DeStefano*.[65] This subsection will also show that even something as seemingly simple as a paper test still fits within the cybernetic framework as a machine. There is no need for a machine to be filled with artificial intelligence or big data to be part of a complex, interdependent cybernetic system.

In *Ricci v. DeStefano*, the City of New Haven refused to certify the results of their supposedly "neutral test" developed by the company, Industrial/Organizational Solutions, Inc. (IOS), for firefighter promotions because the test results exacerbated racial disparities, "ensur[ing] that virtually all of the open promotional positions would have gone to whites."[66] When New Haven declined to certify the results for fear of a disparate impact discrimination lawsuit, the white firefighters (including Frank Ricci) sued New Haven for disparate treatment. The Majority ruled against New Haven because "the record makes clear there is no support for the conclusion that respondents had an objective, strong basis in evidence to find the tests inadequate, with some consequent disparate-impact liability in violation of Title VII."[67] The Majority claimed that New

[63] Elizabeth Fleming & Amy R Pritchett, *SRK as a framework for the development of training for effective interaction with multi-level automation*, 18 COGN. TECHNOL. WORK 511, 513 (2016). (internal citations omitted)

[64] Feigh and Pritchett, *supra* note 47.

[65] *Ricci v. DeStefano*, 557 U.S. 557.

[66] Harris and West-Faulcon, *supra* note 58 at 109.

[67] *Ricci*, 557 U.S. at 585.

Haven was only relying on a "a threshold showing of a significant statistical disparity, and nothing more."[68]

Justice Ginsberg dissented, arguing that Majority "rest[ed] on the false premise that respondents showed 'a significant statistical disparity,' but 'nothing more.'"[69] While Ginsberg's dissent listed numerous concerns, the cybernetic system framework shows how she and the City of New Haven could have better organized and articulated their concerns with what was a cybernetic system. She quoted officials who said "even if individual exam questions had no intrinsic bias, the selection process as a whole may nevertheless have been deficient. The officials urged the CSB to consult with experts about the 'larger picture.'"[70] The cybernetic system framework shows that larger picture.

The following walks through each element of the cybernetic system to show how it could have contributed to the discriminatory output. Within the cybernetic system framework, the firefighter promotion test designed by IOS is the machine (analogous to hiring software), and the user is New Haven (analogous to a hiring manager).

First, the outputs were inadequately defined. IOS was focused on developing a test that was facially neutral,[71] and said any disparate impact was due to external factors, though none were specified.[72] The Majority, being focused on process and anticlassification instead of outcomes and antisubordination sees no issue with this goal because the "questions were relevant"[73] and exams "appea[r] to be… reasonably good."[74] The Majority positively quoted two experts to effectively say that disparities were essentially inevitable. The first expert, Professor Janet Helms,

> "concluded that because 67 percent of the respondents to the job-analysis questionnaires were white, the test questions might have favored white candidates, because 'most of the literature on firefighters shows that the different groups perform the job differently.' Helms closed by stating that no matter what test the City had administered, it would have revealed 'a disparity between blacks and whites, Hispanics and whites,' particularly on a written test."[75]

The second expert, Christopher Hornick, an industrial/organizational psychologist, explained that

> "adverse impact in standardized testing 'has been in existence since the beginning of testing,' and that the disparity in New Haven's test results was 'somewhat higher but generally in the range that we've seen professionally.' He told the CSB he was 'not suggesting' that IOS 'somehow created a test that had adverse impacts that it should not have had.'"[76]

[68] *Ricci*, 557 U.S. at 587.
[69] *Ricci*, 557 U.S. at 644.
[70] *Ricci*, 557 U.S. at 615.
[71] *Ricci*, 557 U.S. at 569.
[72] *Ricci*, 557 U.S. at 567.
[73] *Ricci*, 557 U.S. at 588.
[74] *Ricci*, 557 U.S. at 588.
[75] Ricci, 557 U.S. at 572. (internal citations omitted)
[76] *Ricci*, 557 U.S. at 591 (internal citations omitted).

The problem is that New Haven was concerned about the outputs, not process. New Haven had no interest in a test that produced racial disparities 'but were generally in the range of normal' or a test that would inevitably favor white firefighters. New Haven wanted a test that had no serious racial disparities. This was a city with a long history of racial discrimination in firefighting[77] including a high-profile lawsuit regarding discriminatory practices in New Haven's firefighter hiring that inspired the contract guiding the design of the exam at issue in *Ricci*.[78] Despite New Haven's clear antisubordination focus on avoiding discriminatory outcomes, the Majority, as good adherents to anticlassification, believed that as long as the process was good enough, the discriminatory outcomes did not really matter.

Second, the inputs were inadequate because there are predictable disparities between races on written tests,[79] and most of the analyses used to prepare the firefighter promotion test was based on information gathered from white firefighters.[80]

Third, the test (or machine) itself was inadequate as the third-party designers, IOS,[81] had never designed a promotion examination before.[82] Yes, IOS used job analyses, ride-alongs, questionnaires, and oversampled minorities[83] but that does not mean they used that information appropriately. The questions were not germane to New Haven,[84] and may have been too focused on certain aspects of the job that disadvantaged people who had not been trained in that.[85] The test was mostly memorization from study materials instead of problems firefighters would learn from experience on the ground,[86] and there was unequal access to study materials due to cost and delivery schedule.[87]

Fourth, the human-machine interaction was inadequate because the test was not designed to be suitably precise for ordered ranking or a pass-fail threshold despite that being exactly what New Haven needed.[88]

Fifth, the user was inadequate because while there were 30 external assessors[89] with panels each including one white, one black, and one Hispanic member,[90] New Haven did not request a

[77] *See Ricci*, 557 U.S. at 608-18, showing "the long history of rank discrimination against African–Americans in the firefighting profession," *Ricci*, 557 U.S. at 630 n. 8.

[78] *Firebird Soc. of New Haven, Inc. v. New Haven Bd. of Fire Comm'rs*, 66 F.R.D. 457, 460 (D. Conn.), *aff'd sub nom. Firebird Soc. v. Members of Bd. of Fire Comm'rs, City of New Haven*, 515 F.2d 504 (2d Cir. 1975) (where plaintiffs showed that "the minority population of New Haven was 30 percent, and that none of the 502 men employed by the Department was Hispanic and 18, or less than 4 percent, were black" and "of the 107 officers in the Department only one was black, and he held the lowest rank above private.")

[79] *Ricci*, 557 U.S. at 572.

[80] *Ricci*, 557 U.S. at 617.

[81] *Ricci*, 557 U.S. at 564.

[82] *Ricci*, 557 U.S. at 569.

[83] *Ricci*, 557 U.S. at 564-65.

[84] *Ricci*, 557 U.S. at 613.

[85] *Ricci*, 557 U.S. at 617.

[86] *Ricci*, 557 U.S. at 617-18.

[87] *Ricci*, 557 U.S. at 613-14, 17.

[88] *Ricci*, 557 U.S. at 637 n. 16.

[89] *Ricci*, 557 U.S. at 565-66.

[90] *Ricci*, 557 U.S. at 569.

technical report to better understand the test it had ordered prior to deploying the test.[91]

Sixth, the feedback was inadequate because the New Haven's union contracts required a high weighting of the written test over the oral test[92] which had historically produced disparities in other cities.[93] New Haven also did not use an assessment center which was known to produce more accurate results.[94]

And ultimately, the organizational factors were inadequate because IOS was not allowed to show the exams to anyone prior to administration[95] such that the New Haven officials were unable to check the content to see if it was relevant.[96]

This section shows how cybernetic system discrimination can occur despite good intentions and "good enough" process – a typica of discrimination that anticlassification adherents refuse to see. The framework also shows how to organize and search for sources of discrimination inside each element while ultimately showing how there is an intense interaction between user, machines, organizations far more complex and interdependent then anticlassification's adherents want to accept.

III. THE INTERDEPENDENCE AND COMPLEXITY OF CYBERNETIC SYSTEM DISCRIMINATION

Our legal system, especially antidiscrimination law, believes in three fundamental myths incompatible with cybernetic systems that ultimately undermine its ability to adjudicate cases:[97] (1) deconstruction, that with enough access, discovery, questions and witnesses, the court can understand how a system failed; (2) dualism, that there is usually a meaningful distinction between human contributions to failure and machine contributions to failure; and, (3) structuralism, that every system is made up of specific, articulable, defined building blocks or elements with specific, articulable, defined relationships. The myths of deconstruction, dualism, and structuralism have long discarded by those specializing is system safety and accident investigation and replaced with reality that cybernetic systems are interdependent and complex: the numerous components of a cybernetic system depend upon each other for success[98] such that the specific cause of cybernetic system's failure is incredibly difficult to identify.[99] Ultimately, the law must face more than the consequences of the end of deconstruction and duality. Complex, interdependent cybernetic systems often cause emergent behavior that often drifts toward failure without a clear idea of who is accountable. The law has a long way to go.

A. The Myths of Deconstruction, Dualism, and Structuralism

[91] *Ricci*, 557 U.S. at 566. IOS said it would not add anything. *Ricci*, 557 U.S. at 638.

[92] *Ricci*, 557 U.S. at 570.

[93] *Ricci*, 557 U.S. at 614.

[94] *Ricci*, 557 U.S. at 611-12.

[95] *Ricci*, 557 U.S. at 570, 637.

[96] *Ricci*, 557 U.S. at 616.

[97] DEKKER, *supra* note 52 at 2–4.

[98] Johnson et al., *supra* note 54 at 47. ("Interdependence" describes the set of complementary relationships that two or more parties rely on to manage required (hard) or opportunistic (soft) dependencies in joint activity.)

[99] VICENTE, *supra* note 55 at 14.,

Sidney Dekker introduced three characteristics of the "increasingly obsolete technical worldview" of how we understand cybernetic accidents: deconstruction, dualism, and structuralism.[100] Each of these three plague the legal community. First, deconstruction is the belief "that a system's functioning can be understood exhaustively by studying the arrangement and interaction of its constituent parts."[101] This reverse engineering is seen as the height of success in accident investigations but also legal proceedings. Almost every part of litigation—discovery, direct and cross-examination of witnesses—is built around the idea that if we only had enough access and discovery, asked enough questions, brought in enough witnesses, we could understand what happened. Looking at the cybernetic framework one could theoretically say, "Ok, all we need to know is the inputs, the machine, the human-machine interaction, the user, the feedback, and the organization. Then we will understand exactly what happened." However, that is the wrong takeaway. Cybernetic systems are innately characterized by interdependence and complexity which is in deep conflict with deconstruction, revealing it to be a myth.[102] Accident investigators no longer believe in deconstruction. Instead, they now ask, "What happens if no amount of analysis of the constituent parts can conclusively understand what caused the failure?"

This question is best evidenced by two of the most infamous aviation accidents: TWA 800 and Air France 447. The Boeing 747 flight TWA 800 exploded in midair after takeoff from New York's John F. Kennedy airport in 1996, killing all 230 people on board.[103] It was recovered from the floor of the Atlantic Ocean, reconstructed as part of the "largest and most complex [reconstruction] ever undertaken in the history of civil aviation," and now serves as a key part of training accident investigators.[104] Under the deconstruction hypothesis, "with the puzzle as complete as possible, the broken part(s) should eventually get exposed, allowing investigators to pinpoint the source of the explosion."[105] But despite four years of investigation, parallel investigation by the Federal Bureau of Investigation, the most sophisticated reconstruction in the history of civil aviation, computational analysis of 32 separate scenarios, investigators never determined either the source of the ignition or the mechanism of the ignition.[106] "[T]he reconstructed parts refused to account for the behavior of the whole. In such a case, a frightening, uncertain realization creeps into the investigator corps and into industry. A whole failed without a failed part. An accident happened without a cause; no cause–nothing to fix, nothing to fix–it could happen again tomorrow, or today."[107]

[100] DEKKER, *supra* note 52 at 2.

[101] *Id.* at 2.

[102] *See e.g.*, Mike Ananny & Kate Crawford, *Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability*, 20 NEW MEDIA SOC. 973, 981 (2018). ("Engineers of these systems could not precisely say where in the problems were occurring—even though they had total access to the systems' designs and implementations.").

[103] NAT'L TRANSP. SAFETY BD., IN-FLIGHT BREAKUP OVER THE ATLANTIC OCEAN, TRANS WORLD AIRLINES FLIGHT 800, BOEING 747-131, N93119, NEAR EAST MORICHES, NEW YORK, JULY 17, 1996 1 (2000) [herinafter NTSB TWA 800 REPORT], *available at* https://www.ntsb.gov/investigations/AccidentReports/Reports/AAR0003.pdf.

[104] *NTSB Training Center: Campus*, NATIONAL TRANSPORTATION SAFETY BOARD (accessed May 28, 2021) https://www.ntsb.gov/Training_Center/Pages/facilityloc.aspx#wreckage . *See also*, NTSB TWA 800 REPORT at 139 n. 280.

[105] DEKKER, *supra* note 52 at 2.

[106] NTSB TWA 800 REPORT at 293-94 ("neither the energy release mechanism nor the location of the ignition inside the [central wing tank] could be determined from the available evidence.")

[107] DEKKER, *supra* note 52 at 2.

In 2009, the Airbus A330 flight Air France 447 from Rio de Janeiro crashed two hours after takeoff into the Atlantic Ocean at vertical speed of 124 mph killing all 228 people on board.[108] Ice had formed on key airspeed indicators, resulting in incorrect airspeed readings and the autopilot disconnecting in ways that the pilots did not understand.[109] With the pilots completely disoriented by the alarms and error warnings, Flight 447 crashed into the Atlantic Ocean. The entire accident sequence lasted only four minutes and twenty seconds. The French Bureau of Enquiry and Analysis for Civil Aviation Safety (BEA) was tasked with investigating the accident. After three years, reviewing all 1,300 parameters of flight data, two hours of voice recordings, even setting up a "new working group dedicated to Human Factors made up of pilots…, a specialist in cognitive sciences, a doctor, and BEA investigators," BAE never fully identified what caused the pilots so much confusion when interacting with their autopilot system that they could not save the plane from such a catastrophic end. Their uncertainty (emphasized below) is evident in their language below summarizing the accident trajectory regarding just one of the human-machine interaction factors:

> "The events can be explained by a *combination of the following factors* [including] the crew not taking into account the stall warning, *which could have been due to* (1) A failure to identify the aural warning…, (2) [t]he appearance at the beginning of the event of transient warnings that *could be considered* as spurious, (3) [t]he absence of any visual information to confirm the approach-to-stall after the loss of the limit speeds, (4) [t]he *possible confusion* with an overspeed situation in which buffet is also considered as a symptom, (5) Flight Director indications that *may have led the crew to believe* that their actions were appropriate, even though they were not, [*or*] (6) [t]he difficulty in recognizing and understanding the implications of a reconfiguration in alternate law with no angle of attack protection."[110]

The language shows that the accident in many ways is still unexplained. Imagine attempting to determine who was responsible for the accident when there were a "combination of factors" contributing to the accident, including the users not responding appropriately to a warning "which could have due to" various events like warnings that "could be considered as spurious," "possible confusion" about the situation, or a machine that "may have led the [users] to believe" their actions were appropriate even though they were not. Three years of intense investigations studying what was in effect a 4 minute 20 second human-machine interaction with as complete information as possible still remains unexplained.

Air France 447 also highlights and rejects the second myth plaguing the legal community: dualism. Dualism is the belief that there is a distinct separation between machine causes and human causes, between human error and machine failure – and by extension a further separation from the sociotechnical system they operate within.[111] Again, looking at the cybernetic framework one could say, "yes, the machine and the human are distinct." Again, that would be the wrong

[108] BUREAU OF ENQUIRY AND ANALYSIS FOR CIVIL AVIATION SAFETY, ON THE ACCIDENT ON 1ST JUNE 2009 TO THE AIRBUS A330-203 REGISTERED F-GZCP OPERATED BY AIR FRANCE FLIGHT AF 447 RIO DE JANEIRO – PARIS 24 (2011), [herinafter BEA AF 447 REPORT] *available at* https://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf.

[109] *Id.* at 17.

[110] *Id.* at 200. (emphasis added)

[111] DEKKER, *supra* note 52 at 2–3.

takeaway. The elements are inextricably linked through human-machine interaction and their presence within a sociotechnical system. "All human activity takes place within and is influenced by the environment, both physical and social, in which it takes place."[112]

In discrimination law, as in much of the history of accident investigation, the drive is to find human error, human discrimination. But human error is almost always suspended, unstably, somewhere between the human, the machine, and the sociotechnical system.[113] The error is neither fully human, nor fully machine, nor even fully sociotechnical. Even when it is supposedly found, how do we distinguish isolated human error from the reality that machine and sociotechnical failures will express themselves in the human action. If there is confusion between the pilot, the autopilot, and the procedures in Air France 447 or hiring manager, hiring software, and the hiring procedures, then what is the cause? Human error? Machine error? Sociotechnical error? You need all three to succeed and all three to fail. Where one ends and the others begin is no longer clear. The scientists and engineers investigating accidents know that machines, humans, and their sociotechnical systems are intertwined in ways that resist the neat, dualist, deconstructed disentanglement still favored by discrimination law today.[114] The "cause" of an accident is often determined by how far and how broadly we are willing to look.[115]

The last myth, structuralism, represents our overreliance on the models and language such that we forget to appreciate the "organic, ecological adaptability" of cybernetic systems.[116] Just like the legal community, the system safety community has accepted certain structures, language and models that are used for quantifying, measuring, and modeling failure. These structures are clearly valuable for research and accident investigation but the flaw of structuralism is that too often we forget that these systems of humans, machines, and organizations are not clear, defined, models with actions and relationships that are perfectly discrete and definable. If a certain aspect of a cybernetic system failure does not fit into our accepted model of how failure or discrimination occurs, then we disregard it as if it is irrelevant instead of question if our model is relevant. This is ultimately what creates the cybernetic black holes. As Dekker explained of the system safety and human factors community:

> "[L]anguage, if used unreflectively, easily becomes imprisoning. Language

[112] NANCY LEVESON, ENGINEERING A SAFER WORLD: SYSTEMS THINKING APPLIED TO SAFETY 39 (2011).

[113] This paragraph is adapted from the following to account for sociotechnical systems and anti-discrimination law. DEKKER, *supra* note 52 at 7.

[114] Jens Rasmussen, *Risk management in a dynamic society: a modelling problem*, 27 SAF. SCI. 183–213, 193 (1997). (mapping out how a seemingly simple traffic accident causing an oil spill into a drinking water supply can be "caused" by the interaction of government policy, regulations, local government planning and budgeting, company planning, physical processes of what the driver did, and the equipment and surroundings). J. H. Saleh et al., *Highlights from the literature on accident causation and system safety: Review of major ideas, recent contributions, and challenges*, 95 RELIAB. ENG. SYST. SAF. 1105–1116, 1108 (2010).

[115] See also, Anna Lauren Hoffmann, *Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse*, 22 INF. COMMUN. SOC. 900, 904 (2019). ("Sometimes this means looking at the decisions of specific designers or the demographic composition of engineering or data science teams to identify their social 'blindspots' The idea here is that, as one New York Times article put it, 'software is not free of human influence' because 'algorithms are written and maintained by people' But just as the search for bad actors' places structural issues beyond the law's reach, appealing to the 'blindspots' of particular designers or teams ignores the structuring role of technology, instead reducing a system's shortcomings to the biases of its imperfect human designers.")

[116] DEKKER, *supra* note 52 at 4.

> expresses but also determines what we can see and how we see it. Language constraints how we construct reality. If metaphors encourage us to model accident chains, then we will start our investigation by looking for events that fit that chain. But which events should go in? Where should we start? As Nancy Leveson pointed out, the choice of which events to put in is arbitrary, as are the length, the starting point and level of detail of the chain of events. What, she asked, justifies assuming that initiating events are mutually exclusive, except that it simplifies that mathematics of the failure model."[117]

The structuralism in antidiscrimination is evidenced by the separation between the three worlds of unconscious discrimination, machine discrimination, and systemic discrimination. It shapes the questions we ask in consequential ways. We ask whether the discrimination caused by the machine, the human, or the institution? The question seems normal, simple, and innocent. We investigate, searching for the smoking gun of a discriminatory machine, a discriminatory human, or discriminatory laws and policies. "Looking for failures—human, [machine], or organizational—in order to explain failures is so common-sensical that most investigations never stop to think whether these are indeed the right clues to pursue."[118] We never stop to ask ourselves, what if there is no smoking gun? What if cybernetic system discrimination is just like TWA 800 or Air France 447?

The reason that accident investigators have worked to move beyond the self-constructed limitations of needing to identify a single cause of error, locating the error in either the human or the machine, and overdependence on their models and language, is simply because it better serves their ultimate goal: eliminating accidents altogether. There is only so much that can be done from desperately searching for the lone smoking gun of human error or machine failure. Antidiscrimination law faces the same choice: anticlassification's dedication to process and independent, single sources of discrimination caused by either the human or machine; or antisubordination's dedication to eliminating (or at least, meaningfully addressing) discrimination no matter the process. The following discussion shows how the properties of interdependence and complexity replace the myths of deconstruction, dualism, and structuralism, allowing accident investigators to truly understand and prevent future failures. Our law and policies ought do the same.

B. The Realities of Interdependence and Complexity

1. Interdependence

When humans and machines operate together within an organization, they become interdependent such that what each agent does depends on what each other agent does; requiring coordination in time and space, and some amount of transparency and trust. Formally, interdependence is defined as the complementary relationships to manage required and opportunistic dependencies in joint activity.[119] Dependencies exist when agents lack the required

[117] *Id.* at 4. (citing NANCY G. LEVESON, SYSTEM SAFETY ENGINEERING: BACK TO THE FUTURE (2002).)

[118] DEKKER, *supra* note 52 at 5.

[119] Johnson et al., *supra* note 54 at 47. ("Interdependence" describes the set of complementary relationships that two or more parties rely on to manage required (hard) or opportunistic (soft) dependencies in joint activity.)

capacity (e.g., knowledge, skills, abilities, or resources) to competently perform an activity individually. [120]

In the context of hiring software, imagine software used to filter thousands of job applicants into a few dozen for the hiring manager to review. The hiring manager does not have the required time to review thousands of job applicants and therefore, relies on the software to select a subset of applications. The software relies on the hiring manager to identify what types of applications are preferred. Both the hiring manager and the software rely on each other to complete the joint activity of effectively selecting a few dozen applicants to interview from thousands. A successful applicant must have satisfied both the machine screening and the human screening. Both the software and the hiring manager operate within an organization who purchased the software, trained the manager, and controls the manager's resources, their time and processes.

But beyond the fact that the hiring manager is now reliant upon the software to achieve her goal and vice versa, it is critical to understand that the hiring manager is no longer hiring on her own; she is now performing a joint activity. "[A] person's processes may be very different in individual and joint actions, even when they appear identical."[121] In these joint activities, unlike individual activities, users are constantly asking and answering questions such as "What information needs to be shared?," "Who needs to share the information with whom?," and "When is the information relevant to share?"[122] Canonical examples of individual versus joint activities include musical solos versus duets, running alone versus running a relay race, driving alone versus driving in traffic or leading a caravan of vehicles.[123] But we experience the difference between individual and joint activities every day of our lives: entering through a door on our own is different than entering a door with others behind us, or in legal practice writing a brief alone is different than writing it jointly. Using different language, joint activity is said requires coordination and teamwork.[124]

Specifically, in the context of human-machine interaction, by adding the software into the hiring manager's workflow, and requiring coordination and teamwork, her work is fundamentally different now. "Adding or expanding the role of automation changes the nature of the interactions in the system, often affecting the humans' role in profound ways."[125] Users are often left perplexed, asking familiar questions we have asked of our phones and computers endless times, "what is it

[120] "Dependence exists when an entity lacks a required capacity to competently perform an activity in a given context." "Capacity is the total set of inherent things (e.g., knowledge, skills, abilities, and resources) that an entity requires to competently perform an activity individually." *Id.* at 47. *See also,* Matthew Johnson et al., *The fundamental principle of coactive design: Interdependence must shape autonomy*, *in* COORDINATION, ORGANIZATIONS, INSTITUTIONS, AND NORMS IN AGENT SYSTEMS VI 172–191 (2011); Johnson et al., *supra* note 54. P. J. Feltovich et al., *Toward an Ontology of Regulation: Socially-Based Support for Coordination in Human and Machine Joint Activity, in* PRE-PROCEEDINGS OF THE ENGINEERING SOCIETIES IN THE AGENT'S WORLD 06 (ESAW06) (2006).

[121] Johnson et al., *supra* note 47 at 51. (quoting HERBERT H. CLARK, USING LANGUAGE (1996).)

[122] Johnson et al., *supra* note 54 at 51–52.

[123] Gary Klein et al., *Common Ground and Coordination in Joint Activity, in* ORGANIZATIONAL SIMULATION 139 (William B. Rouse & Kenneth R. Boff eds., 2005). (summarizing examples)

[124] G Klien et al., *Ten challenges for making automation a "team player" in joint human-agent activity,* 19 IEEE INTELL. SYST. 91 (2004).

[125] Klaus Christoffersen & David D. Woods, *How to Make Automated Systems Team Players*, 2 *in* ADVANCES IN HUMAN PERFORMANCE AND COGNITIVE ENGINEERING RESEARCH: AUTOMATION 1, 3 (Eduardo Salas ed., 1. ed ed. 2002).

doing? Why is it doing that? What is it going to do next?"[126] "It is the joint nature of key tasks that defines the heart of collaborative activity—and it is the effective management of interdependence that makes such work possible. Therefore, effective management of systems with autonomy requires an understanding of the impact a change in autonomy may have on the interdependence in the human-machine system."[127]

And if the user's work is now fundamentally different, then the mechanisms of failure are fundamentally different, too. It is widely understood that there are at least seven ways that the introduction of machines into human workflow can result in "deadly" outcomes:[128] (1) the work is transformed, users' roles change (*the envisioned world problem*);[129] (2) users are required to perform new kinds of work, and do it more, do it faster, or in more complex ways (*the law of stretched systems*); (3) users are required to track more sets of information making it more difficult to remain aware of and integrate all the new activities and changes (*continuous coordination costs*); (4) new knowledge and skill demands are imposed on the user and the human may not have sufficient context to make decisions because they are practically left out of the loop (*automation surprises*); (5) new levels and types of feedback are needed to support new roles (*continuous coordination costs*); (6) resulting explosion of flexibility, features, options, and modes creates new demands, types of errors, and paths towards failure (*automation surprises*); and (7) both machines and users are fallible and machines may obscure the information necessary for user decision making (*principles of complexity*).

The use of "deadly" outcomes here is not hyperbole. One of the foundational sources of system safety and accident investigations, human factors, and cognitive engineering research is aviation.[130] Aviation focuses so much on these failures because its history is in many ways best understood through high-profile catastrophes like TWA 800 and Air France 447. Observers describe the industry as constantly at "war against the death and destruction caused by… aviation accidents."[131] Those of us in the aviation community are even more explicit. Mary Schiavo, former Inspector General of the U.S. Department of Transportation, stated that with respect to aviation, "[w]e regulate by tombstones."[132] We call it "tombstone design" where rules, procedures, and regulations are only updated once enough people have died – they "are written in blood."[133] In 1989, in response to high-profile fatal accidents, the Air Transport Association of America (ATA)

[126] *Id.* at 3. (citing EARL L WIENER, *Human factors of advanced technology ("Glass Cockpit") transport aircraft.* (1989).)

[127] Matthew Johnson, *Coactive Design: Designing Support for Interdependence in Human-Robot Teamwork*, 2014.

[128] Jeffrey M. Bradshaw et al., *The Seven Deadly Myths of "Autonomous Systems"*, 13 IEEE INTELL. SYST. 2–9, 7 (2013). (summarizing Table 1).

[129] Miller and Feigh, *supra* note 47.

[130] Canellas and Haga, *supra* note 47 at 31. (Aviation is the "canonical domain for understanding human-automation interaction in complex, safety-critical operations.") Even the National Highway Transportation Safety Administration, when discussing rules and regulations for autonomous vehicles, looks to the aviation domain, saying that the "lessons learned through the aviation industry's experience with the introduction of automated systems may be instructive and inform the development of thoughtful, balanced approaches." AUTOMATED VEHICLES 3.0: PREPARING FOR THE FUTURE OF TRANSPORTATION, 42 (2018).

[131] MARK HANSEN, CAROLYN MCANDREWS & EMILY BERKELEY, *History of Aviation Oversight in the United States* (2005).

[132] MARY SCHIAVO, FLYING BLIND, FLYING SAFE 65 (1997).

[133] Rod Rakic, *Regulations are Written in Blood: Why Planesharing is Grounded for Now*, AIRCRAFT OWNERS AND PILOTS ASSOCIATION BLOG (August 7, 2014), https://blog.aopa.org/aopa/2014/08/07/planesharing/.

established a task force to examine the impact of automation on aviation safety. The legal community should heed their prescient conclusion:[134]

> "During the 1970s and early 1980s... the concept of automating as much as possible was considered appropriate. The expected benefits were a reduction in pilot workload and increased safety... Although many of these benefits have been realized, serious questions have arisen and incidents/accidents have occurred which question the underlying assumption that maximum available automation is always appropriate or that we understand how to design automated systems so that they are fully compatible with the capabilities and limitations of the humans in the system."

2. Complexity

The interdependence between humans, machines, and the organizations they operate within, reveal the second characteristic innate to cybernetic systems: complexity. Humans interact with their organizational structures (the social system) and utilize their machines (the technical system) to achieve overall system goals and objectives.[135] "These systems involve context-rich workplace settings, organizational structure, human operators, and sophisticated technology that when taken collectively are known as complex sociotechnical systems."[136]

Complexity typically means a system can be characterized by some of the following features: (a) large problem spaces and ill-structured problems; (b) mediated interaction and automation; (c) heterogeneous, shifting, ill-defined, or competing goals among multiple or distributed agents; (d) high stakes or time stressed; and (e) dynamic, coupled action and feedback loops in an uncertain environment.[137] "What makes sociotechnical systems complex is the simple fact that all [features] are simultaneously at play when considering the performance of complex systems. Therefore, [we must] consider[] how all factors present themselves in the system of interest, in order to understand how they collectively influence the behaviors of the system."[138]

To describe these factors, I once again use *Ricci*. Even though the machine in that case is a paper test, and far from what we consider automation or artificial intelligence, it is still a machine for the purposes of cybernetic system. The takeaway is that these interactions, even with a paper test, are far more complex than what many commentators, courts, and the majority in *Ricci* want

[134] CHARLES E BILLINGS, AVIATION AUTOMATION: THE SEARCH FOR A HUMAN-CENTERED APPROACH (Lawrence Erlbaum ed., 1997).

[135] Marc C. Canellas et al., *Framing Human-Automation Regulation: A New Modus Operandi from Cognitive Engineering*, *in* WEROBOT 2017, 6 (2017).

[136] Marc C. Canellas et al., *Framing Human-Automation Regulation: A New Modus Operandi from Cognitive Engineering*, *in* WEROBOT 2017, 6 (2017). For complex sociotechnical systems, *see generally,* Baxter and Sommerville, *supra* note 45; Walker et al., *supra* note 45; Waterson et al., *supra* note 45. For cognitive systems engineering, *see generally*, Hollnagel and Woods, *supra* note 45; RASMUSSEN, PEJTERSEN, AND GOODSTEIN, *supra* note 45; WOODS AND HOLLNAGEL, *supra* note 45; Woods and Roth, *supra* note 45.

[137] Integrating the attributes of complexity considered by cognitive systems engineering, VICENTE, *supra* note 55 at 14., with the naturalistic decision making factors, Judith Orasanu & Terry Connolly, *The Reinvention of Decision Making*, *in* DECISION MAKING IN ACTION: MODELS AND METHODS 3 (Gary A Klein et al. eds., 1993)., as suggested by Canellas et al., *supra* note 44 at 8.

[138] Canellas et al., *supra* note 44 at 7.

to believe.[139]

a. Large Problem Spaces and Ill-Structured Problems

"Real decision problems rarely present themselves in neat, complete form."[140] These ill-structured problems require the decision maker to do significant work to even recognize the situation and form hypotheses about what is happening. In addition to being ill-structured, real decision problems are often composed of many different elements and forces – humans often face an innumerable range of possibilities that must be dealt with without exceeding their resource limitations.[141]

In *Ricci,* the problem of designing a written exam to determine the best qualified firefighter officers is particularly ill-structured with many elements and possibilities. The designer for the written exam, IOS, interviewed numerous officers, rode with and observed the officers, and wrote and administered job-analysis questionnaires.[142] IOS used all of this to construct a 100-question multiple choice exam written below a 10th-grade reading level to determine who was qualified to be a firefighter officer.[143] But while the majority in *Ricci* believed that effort is evidence of efficacy, to most who have taken or designed multiple choice exams, there is often a gap between what the multiple choice exam is capable of measuring and what we want to measure. Ask a practicing attorney how well the Law School Admissions Test really determined the best qualified law students or how well the Multistate Bar Examination.

As explained by Justice Ginsberg, "[t]hat IOS… may have been diligent in designing the exams says little about the exams' suitability for selecting fire officers."[144] Dr. Christopher Hornick, an industrial and organizational psychology consultant with 25 years' experience with police and firefighter testing explained that he had "never one time ever had anyone in the fire service say to me, 'Well, the person who answers—gets the highest score on a written job knowledge, multiple-guess test makes the best company officer.' We know that it's not as valid as other procedures that exist."[145] According to Justice Ginsburg, "significant doubts had been raised about whether [IOS] properly assessed the key attributes of a successful fire officer. … 'Upon close reading of the exams, the questions themselves would appear to test a candidate's ability to memorize textbooks but not necessarily to identify solutions to real problems on the fire ground.'"[146]

[139] See also, DAVID D. WOODS & ERIK HOLLNAGEL, JOINT COGNITIVE SYSTEMS: PATTERNS IN COGNITIVE SYSTEMS ENGINEERING (2006) back cover. ("Our fascination with new technologies is based on the assumption that more powerful automation will overcome human limitations and make our systems `faster, better, cheaper' resulting in simple, easy tasks for people. But how do new technology and more powerful automation change our work? What [cognitive systems engineering has] found is not stories of simplification through more automation but stories of complexity and adaptation. … Ironically, more autonomous machines have created the requirement for more sophisticated forms of coordination across people, and across people and machines, to adapt to new demands and pressures.")

[140] Orasanu and Connolly, *supra* note 138 at 7.

[141] VICENTE, *supra* note 55.

[142] Ricci, 557 U.S. at 565-66.

[143] Ricci, 557 U.S. at 566.

[144] Ricci, 557 U.S. at 637.

[145] Ricci, 557 U.S. at 616.

[146] Ricci, 557 U.S. at 617-18.

b. Mediated Interaction and Automation

"[I]t is often the case that the goal-relevant properties of a complex sociotechnical system cannot be directly observed by human perceptional systems unaided. … In these cases, it is usually not possible for people to go out and directly gather information using the powerful perceptual systems that serve them so well in the natural environment."[147] In modern cybernetic systems, "[c]omputer algorithms control the work domain, and the workers' responsibility is to monitor the state of the automation and the work domain itself."[148]

For example, how does a city determine who will be the best firefighter officers when these firefighters have never had experience or an opportunity to show their competence as an officer? Metrics, measures, evaluations, and assessments are all used to *approximate* qualities of interest but cannot perfectly replicate and predict performance. Particularly in the context of hiring where allegedly "objective" measures and machines are deliberately placed between the hiring manager and the applicant with the goal of reducing bias. In *Ricci*, New Haven officials were explicitly prohibited from checking the content of the questions prior to their administration, so IOS hired a third-party to review the exams content and fidelity.[149]

Here, and in most cases, the hiring manager is not seeing the applicant in the actual environment of the job or even the applicant while they were performing the assessment. Instead, they are merely seeing a mediating representation, no better than a bar exam score or LSAT score allows a state or law school to see an actual applicant in the context of being an attorney or law student. The organization's goal is to replace the hiring manager's interpersonal skills and experience to determine who is the best qualified and instead rely on the hiring manager's ability to reason through numbers, metrics, and measures on paper. Therefore, when using machines like tests or software, or other mediating systems, new and more complex skills and cognitive resources are needed to get the job done.[150]

c. Heterogeneous, Shifting, Ill-defined, or Competing Goals Among Multiple or Distributed Agents

"[I]t is rare for a decision to be dominated by a single, well-understood goal or value."[151] Complex sociotechnical systems are also often "[c]omposed of many people who must work together to make the overall system function properly… create[ing] a strong need for clear communication to effectively coordinate the actions of the various parties involved."[152] These agents within the complex, sociotechnical system – be they human, machine, or organization – often have different roles, backgrounds, or locations[153] such that decisions are driven by multiple goals, not all of them clear, and potentially in conflict with each other.[154] As observed by Dörner,

[147] VICENTE, *supra* note 55 at 16. (Citing (Vicente & Rasmussen, 1990))
[148] VICENTE, *supra* note 55. (discussing the "automation" factor)
[149] Ricci, 557 U.S. at 614.
[150] VICENTE, *supra* note 55 at 16.
[151] Orasanu and Connolly, *supra* note 138 at 8.
[152] VICENTE, *supra* note 55. (discussing the "social" factor)
[153] *Id.* (discussing the "distributed" factor")
[154] Orasanu and Connolly, *supra* note 138 at 8. VICENTE, *supra* note 55. (discussing Heterogenous)

"Contradictory goals are the rule, not the exception, in complex situations."[155] So while, the organization in which the humans and machines operate may provide the background or higher-level goals guiding the humans and machines through rules, standard operation procedures or other guidelines,[156] many organizations implicitly pass on the responsibility for reconciling these conflicting goals to the individuals or the machine – referred to as shifting responsibility from the blunt end to the sharp end.[157]

In *Ricci*, there are numerous heterogeneous, shifting, ill-defined, and competing individual and organizational goals. New Haven outsourced the development of the exam process to IOS but was New Haven was legally and socially responsible for the outcome of the exam. So, while IOS was worried about making the test "facially neutral,"[158] and the questions "relevant"[159] and "[faithful] to the source material,"[160] New Haven had to be concerned the potentially discriminatory impact of the exams and any related liability. While New Haven ultimately needed the exam to produce a strict rank ordering of the officer candidates and a clear threshold differentiating pass and fail, IOS did not design the test to those specifications, potentially precluding New Haven from succeeding in showing the potential for discriminatory impact.[161]

Moreover, New Haven tried to adhere to its two-decades-old contract with the local firefighters' union to use combination of written exam and oral exam weighted at 60% and 40% respectively, and contracted IOS to create that exam, even though the contract requirements likely undermined the validity of the exam.[162] IOS was aware that "alternative methods might better measure the qualities of a successful fire offer" but explained that they kept to the outdated test design[163] because of the contract.[164]

d. High Stakes or Time Stress

Complex, sociotechnical systems often operate in environments where failure can have catastrophic consequences to things like life, liberty, and money. Because the outcomes are of real

[155] DIETRICH DORNER, THE LOGIC OF FAILURE: RECOGNIZING AND AVOIDING ERROR IN COMPLEX SITUATIONS. 65 (1989).

[156] Orasanu and Connolly, *supra* note 138 at 10.

[157] Sidney Dekker & Shawn Pruchnicki, *Drifting into failure: theorising the dynamics of disaster incubation*, 15 THEOR. ISSUES ERGON. SCI. 534, 537 (2013). (citations omitted)

[158] *Ricci* 557 U.S. at 569 (quoting the representative from IOS discussing the exam).

[159] *Ricci* 557 U.S. at 571 (quoting a representative from the U.S. Department of Homeland Security discussing the exam).

[160] *Ricci* 557 U.S. at 614-15 (quoting a representative from IOS discussing the responsibilities of the third-party they retained to review the exams).

[161] *Ricci* 557 U.S. at 637 n. 16.

[162] *Ricci* 557 U.S. at 611.

[163] *Ricci* 557 U.S. at 635 ("Testimony before the [New Haven board reviewing the test results] indicated that these alternative methods were both more reliable and notably less discriminatory in operation. According to Donald Day of the International Association of Black Professional Firefighters, nearby Bridgeport saw less skewed results after switching to a selection process that placed primary weight on an oral exam. And Hornick described assessment centers as 'demonstrat[ing] dramatically less adverse impacts' than written exams.") (internal citations omitted)

[164] *Ricci* 557 U.S. at 611-12. *Ricci* 557 U.S. at 637 ("IOS worked within the City's constraints. [IOS] never discussed with the City the propriety of the 60/40 weighting and 'was not asked to consider the possibility of an assessment center.' The IOS exams, [IOS] admitted, had not even attempted to assess 'command presence': '[Y]ou would probably be better off with an assessment center if you cared to measure that.'") (internal citations omitted).

significance to the players involved,[165] decision makers "cannot rely on trial-and-error approaches. … There is a very strong requirement to 'get it right the first time.'"[166] These stakes, especially if the system is operating in an unexpected or potentially problematic way, often create time stress where decisions must be made in timelines faster than would be ideal or even normal.[167] The high stakes and time stress can cause decision makers to feel high levels of personal stress, potentially exhaustion, and loss of vigilance.

Ricci was a study of high stakes decision making. The majority focused on the plaintiff firefighters, emphasizing that the promotion exams "were infrequent, so the stakes were high. The results would determine which firefighters would be considered for promotions during the next two years, and the order in which they would be considered. Many firefighters studied for months, at considerable personal and financial cost."[168] To the majority, the high stakes made New Haven's decision "all the more severe."[169]

Conversely, Justice Ginsberg and New Haven emphasized that the City was facing high stakes, too. There is a long history of racial discrimination in firefighting[170] including a high-profile lawsuit regarding discriminatory practices in New Haven's firefighter hiring that inspired the contract guiding the design of the exam at issue in *Ricci*.[171] Once the exam was started, even if the outcomes had a potentially disparate impact, "changing the weighting formula… could well have violated Title VII's prohibition of altering test scores on the basis of race."[172] Nor was New Haven legally permitted to band the results to make the minority test scores appear higher.[173] At least the majority gave the City this much credit, "Confronted with arguments both for and against certifying the test results—and threats of a lawsuit either way—the City was required to make a difficult inquiry."[174]

e. Dynamic, Coupled Action and Feedback Loops in an Uncertain Environment

[165] Orasanu and Connolly, *supra* note 138 at 9–10.

[166] VICENTE, *supra* note 55. (discussing the "hazard" factor)

[167] Orasanu and Connolly, *supra* note 138 at 9. (discussing the "high stakes" factor)

[168] *Ricci*, 557 U.S. at 562.

[169] *Ricci*, 557 U.S. at 593.

[170] *See Ricci*, 557 U.S. at 608-18, showing "the long history of rank discrimination against African–Americans in the firefighting profession," *Ricci*, 557 U.S. at 630 n. 8.

[171] *Firebird Soc. of New Haven, Inc. v. New Haven Bd. of Fire Comm'rs*, 66 F.R.D. 457, 460 (D. Conn.), *aff'd sub nom. Firebird Soc. v. Members of Bd. of Fire Comm'rs, City of New Haven*, 515 F.2d 504 (2d Cir. 1975) (where plaintiffs showed that "the minority population of New Haven was 30 percent, and that none of the 502 men employed by the Department was Hispanic and 18, or less than 4 percent, were black" and "of the 107 officers in the Department only one was black, and he held the lowest rank above private.")

[172] *Ricci*, 557 U.S. 589-90.

[173] *Ricci*, 557 U.S. 590 ("A state court's prohibition of banding, as a matter of municipal law under the charter, may not eliminate banding as a valid alternative under Title VII. See 42 U.S.C. § 2000e–7. We need not resolve that point, however. Here, banding was not a valid alternative for this reason: Had the City reviewed the exam results and then adopted banding to make the minority test scores appear higher, it would have violated Title VII's prohibition of adjusting test results on the basis of race. § 2000e–2(l); *see also, Chicago Firefighters Local 2 v. Chicago*, 249 F.3d 649, 656 (C.A.7 2001) (Posner, J.) ('We have no doubt that if banding were adopted in order to make lower black scores seem higher, it would indeed be ... forbidden'). As a matter of law, banding was not an alternative available to the City when it was considering whether to certify the examination results.")

[174] *Ricci*, 557 U.S. 593.

Participants in complex sociotechnical systems have actions and feedback loops such that performance or failure is not a single event but series of events.[175] These coupled interactions between participants change over time such that (1) there is a delay between actions and the effect of those actions so participants have to anticipate and act well before the effects are truly known,[176] and (2) it is "very difficult to predict all of the effects of an action, or to trace all of the implications of a disturbance because there are many, perhaps diverging, propagation paths."[177] These dynamic, coupled interactions mean that the user operates in an uncertain environment[178] populated by unanticipated events.[179] Uncertainty is inherent to complex sociotechnical systems, especially due to mediated interaction, such that "the true state of the [system] is never known with perfect certainty."[180] Users must continually "go beyond the information given"[181] to distinguish between spurious suggestions of failure versus true evidence of failure. Assuming the user even knows they are facing an unanticipated event, the user must "improvise and adapt" based on just a "conceptual understanding" of the system because the "normal work procedures no longer apply."[182] In fact, the advice for designers of complex sociotechnical systems is that "design cannot be based solely on expected or frequently encountered situations. … Instead [they] must also operate effectively even – or especially – under idiosyncratic rare events that are not anticipated by workers or designers."[183]

The analysis of *Ricci* through the cybernetic framework in Sec. II.B. shows how many dynamic, coupled actions and feedback loops there are. However, New Haven was largely prohibited from altering the test results, and therefore must "consider[], before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race."[184] Attempting to predict ex ante all the implications of a dynamic, coupled system with action and feedback loops especially when there is a delay between actions and effects, is a nearly impossible task.

C. Consequences of Complex, Interdependent Cybernetic Systems

This Essay has shown that cybernetic systems are all around us, including tests and modern software[185] and that cybernetic systems are characterized by interdependence and complexity,[186] which are in direct opposition to the outdated myths of deconstruction, duality, and

[175] Orasanu and Connolly, *supra* note 138 at 9. (discussing the "action/feedback loops" factor)

[176] VICENTE, *supra* note 55 at 15. (discussing the "dynamic" factor)

[177] "For example, if a particular action is selected to affect goal X, workers must also consider whether that same action will also affect goals Y and Z. These other effects may not be desirable so workers must consider them before acting. Reasoning in a highly coupled work domain puts a great burden on workers because of all the factors that need to be considered at the same time." VICENTE, *supra* note 55. (discussing the "coupled" factor)

[178] *Id.* at 16. (discussing the "uncertainty" factor). Orasanu and Connolly, *supra* note 138 at 8. (discussing the "uncertain dynamic environments").

[179] VICENTE, *supra* note 55 at 16–17. (discussing the "disturbances" factor)

[180] *Id.* at 16.

[181] *Id.* at 16.

[182] *Id.* at 16.

[183] *Id.* at 17.

[184] *Ricci*, 557 U.S. at 585.

[185] Sec. II.

[186] Sec. III.B.

structuralism.[187] What are the consequences? The rest of Essay paper explains the consequences for the enforcement of anti-discrimination law in the United States. However, here, I walk through three high-level consequences that inform all cybernetic systems and therefore the rest of the paper. In short, complex, interdependent systems often causes emergent behavior that often drifts toward failure without a clear idea of who is accountable.

1. The Behavior is Emergent

Emergence occurs when "components within a system… interact to cause outputs or states which cannot be predicted by accounting for the deterministic behavior of the components"[188] Emergence is reality's counter to deconstruction's mythology that the performance of the whole can be predicted by the performance of the parts. Especially when adding in machines and automation, "instead of eliminating error, automation… generates new types of error arising from problematic human-automation interaction."[189] Engineers and accident investigators have long known that "we are building systems that spend more time in nominal operation (that is, are generally better behaved) than previous generations, but when they do operate off nominal, are much further from the nominal than previous generations."[190] In other words, cybernetic systems may rarely fail, but when they do, they failure surprisingly and spectacularly.[191] And as exemplified by TWA 800 and Air France 447,[192] despite the wide awareness and desire to understand, model, simulate, or predict emergent behavior, in many ways even experts are still wandering in the dark.

2. The System Can Drift Toward Failure

Drift is a key emergent behavior of cybernetic systems.[193] Driven by interdependence and complexity, the drift toward failure often progresses in steps so small that they are hardly noticed against a background of dynamic and uncertain humans and machines interacting within sociotechnical systems. Drifting into failure is not always "about breakdowns or malfunctioning of components, as it is [often] about an organization not adapting effectively to cope with the complexity of its own structure and environment."[194] "[S]uccessful outcomes keep giving away the impression that risk is under control" even as the system progresses toward failure.[195] This

[187] Sec. III.A.

[188] Canellas et al., *supra* note 44 at 40. (citing W. Clifton Baldwin, Wilson N. Felder, & Sauser, *Taxonomy of increasingly complex Systems*, 9 INT. J. IND. SYST. ENG. 298–316, 306 (2011); Wilson N. Felder & Paul Collopy, *The elephant in the mist: What we don't know about the design, development, test and management of complex systems*, 1 J. AEROSP. OPER. 317–327, 320 (2012).)

[189] Amy R. Pritchett, *Aviation Automation: General Perspectives and Specific Guidance for the Design of Modes and Alerts*, 5 REV. HUM. FACTORS ERGON. 82–113, 83 (2009). *See also*, Thomas B Sheridan & Raja Parasuraman, *Human-automation interaction*, 1 REV. HUM. FACTORS ERGON. 89–129 (2005); E. L. Wiener & R. E. Curry, *Flight-deck automation: Promises and problems*, 23 ERGONOMICS 995–1011 (1980).

[190] Felder and Collopy, *supra* note 189 at 321.

[191] Canellas et al., *supra* note 44 at 42. (citing Felder and Collopy, *supra* note 189.)

[192] *See also*, Canellas et al., *supra* note 44 at 42. (discussing how the entire aviation industry is still figuring out how predict loss-of-control in-flight and controlled flight into terrain, two of the biggest causes of aviation accidents sharing the same source: human-machine interaction)

[193] Dekker and Pruchnicki, *supra* note 158.

[194] SIDNEY DEKKER, DRIFT INTO FAILURE: FROM HUNTING BROKEN COMPONENTS TO UNDERSTANDING COMPLEX SYSTEMS 121 (2011).

[195] *Id.* at 106.

belies the tension between the goals of efficiency and safety resulting in conflicts that must be negotiated daily by users and organizations. When success is measured by cost, efficiency, *and* the absence of failure, there are many ways that these systems intentionally and unintentionally avoid *seeing* failure in the first place.

Experts and automation are celebrated for their ability to fine-tune their work, compensate for problems and dangers, removing redundancy, eliminating unnecessary expense, and expanding capacity.[196] But as Weick and Sutcliffe explain for both humans and the systems they construct, "[s]uccess narrows perceptions, changes attitudes, reinforces a single way of doing business, breeds overconfidence in the adequacy of current practices, and reduces the acceptance of opposing points of view."[197] Individuals may believe their decisions or influence on the overall performance of a cybernetic system to be so small that they could not possibly cause a failure. The "warning of an incomprehensible and unimaginable event cannot be seen, because it cannot be believed."[198] And as Norman explains, machines and automation are often deployed in a brittle way: taking over control of decision making from people without the ability to handle the off-nominal situations in the way people can.[199] Furthermore, implementation of valuable modern safety and resiliency principles like defense-in-depth can actually contribute to obfuscating the cybernetic system's drift towards failure by concealing the occurrence of hazardous states.[200] All these factors can result in a "normalization of deviance" where "a group's construction of risk can persist even in the face of continued (and worsening) signals of potential danger."[201]

The development of the theories of systemic discrimination are evidence that at least some of the legal community already understand drift to some extent. In challenging the legal understanding of intentional employment discrimination, Linda Hamilton Krieger argued that "biases 'sneak up on' the decisionmaker, distorting bit by bit the data upon which his decision is eventually based."[202] In other words, the decisionmaker drifts towards biased decisionmaking.

3. The Accountability Gap

If the user's work has fundamentally changed from what was likely designed and intended, resulting in unpredictable emergent behavior that drifts subtly towards failure, how are we possibly able to determine who is accountable for the failure? This is the ultimate question in this Essay as applied to our anti-discrimination laws but has long been a question for engineers and accident

[196] William H. Starbuck & Frances J. Milliken, *Challenger: Fine-Tuning the Odds Until Something Breaks*, 25 J. MANAG. STUD. 319, 333 (1988).

[197] KARL E. WEICK & KATHLEEN M. SUTCLIFFE, MANAGING THE UNEXPECTED: RESILIENT PERFORMANCE IN AN AGE OF UNCERTAINTY 52 (Second ed. 2007).

[198] CHARLES PERROW, NORMAL ACCIDENTS: LIVING WITH HIGH-RISK TECHNOLOGIES 23 (1984). Stated differently: "seeing what one believes and not seeing that for which one has no beliefs are central to sensemaking. Warnings of the unbelievable go unheeded." KARL E. WEICK, SENSEMAKING IN ORGANIZATIONS 87 (1995).

[199] Donald A Norman, *The "problem" with automation: inappropriate feedback and interaction, not'over-automation'*, 327 PHILOS. TRANS. R. SOC. B BIOL. SCI. 585–593 (1990).

[200] Francesca M. Favarò & Joseph H. Saleh, *Observability-in-Depth: An Essential Complement to the Defense-in-Depth Safety Strategy in the Nuclear INdustry*, 46 NUCL. ENG. TECHNOL. 803, 804 (2014).

[201] Dekker and Pruchnicki, *supra* note 158 at 5. (summarizing "normalization of deviance" described originally by DIANE VAUGHAN, THE CHALLENGER LAUNCH DECISION: RISKY TECHNOLOGY, CULTURE, AND DEVIANCE AT NASA 394 (1996).)

[202] Krieger, *supra* note 22 at 1188.

investigators[203] of cybernetic systems from nuclear reactors and spacecraft[204] to autonomous weapons[205] and autonomous vehicles.[206] These issues of cybernetic systems are prevalent in cybersecurity, too, where the inability to attribute actions "pose[s] problems for deterrence (because if you cannot identify the perpetrators, you cannot threaten them) and for enforcing the law (because you cannot hold unidentifiable perpetrators accountable)."[207] Cybernetics is a black hole for antidiscrimination law as well. How can the law deter discrimination when plaintiffs cannot find the single, specific causes of the discrimination as the law demands or enforce antidiscrimination when the plaintiffs cannot find the required smoking gun of intent?

It is true that some machine designers, users, or organizations could use the realities of cybernetic systems to *mask* their involvement in the failure.[208] In the context of discrimination, "any form of discrimination that happens unintentionally can also be orchestrated intentionally."[209] Alternatively, they could use *agency laundering*, invoking the complexity or automated nature of an algorithm to explain why the suspect action occurred, allowing them to imply that the action is unintended and something for which they are not responsible.[210] Simply put "blame the machine."

However, there are many accidents where human at the sharp-end is deemed organizationally and legally responsible for an outcome without having sufficient authority to control the outcome.[211] In the engineering community, this is called the authority-responsibility mismatch, although in policy and sociology circles it has been called the "moral crumple zone"[212] or the

[203] Sidney W. A. Dekker, *When human error becomes a crime*, 3 HUM. FACTORS AEROSP. SAF. 83 (2003); Mary L. Cummings, *Automation and Accountability in Decision Support System Interface Design*, 32 J. TECHNOL. STUD. 23 (2006); WORKSHOP DISCUSSION NOTES: ALGORITHMIC ACCOUNTABILITY, 1–5 (2014); David D. Woods, *Conflicts between Learning and Accountability in Patient Safety*, 54 DEPAUL LAW REV. 485 (2005). IEEE, *Ethically Aligned Design: A Vision for Prioritizing Wellbeing with Artificial Intelligence and Autonomous Systems* 236–243 (2019).

[204] JAMES REASON, HUMAN ERROR 173–216 (2009). (describing major accidents including Three Mile Island, Bhopal, Challenger, Chernobyl, Zeebrugge, and the King's Cross underground fire.)

[205] BONNIE LYNN DOCHERTY, MIND THE GAP: THE LACK OF ACCOUNTABILITY FOR KILLER ROBOTS (2015); Vincent C Müller, *Drones and Responsibility: Legal, Philosophical, and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons*1–16 (Ezio Di Nucci & Filippo Santoni de Sio eds., 2016); Marc Canellas & Rachel Haga, *Lost in translation: Building a common language for regulating autonomous weapons*, 35 IEEE TECHNOL. SOC. MAG. 50–58 (2016).

[206] Canellas and Haga, *supra* note 47.

[207] Matthew C. Waxman, *Cyber-Attacks and the Use of Force: Back to the Future of Article 2(4)*, 36 YALE J. INT. LAW 421, 445 (2011).

[208] Johnson, *supra* note 21 at 1019. ("The multiplicity of factors that enter sentencing decisions and the consequent need for discretion may make the inference of race-based decision-making riskier in the sentencing context than where only a few permissible considerations enter into a decision, as Powell argued. However, it also increases the likelihood that race will play a role in the decision: the greater number of factors allows the conscious but covert racist to conceal his or her motives, and the difficulty of weighing all the factors allows the well-intentioned unconscious racist to be influenced—at the margin—by race."

[209] Barocas and Selbst, *supra* note 56 at 692.

[210] Alan Rubel, Adam Pham & Clinton Castro, *Agency Laundering and Algorithmic Decision Systems, in* INFORMATION IN CONTEMPORARY SOCIETY 590–598 (N. Taylor et al. eds., 2019), https://doi.org/10.1007/978-3-030-15742-5_56.

[211] Amy R Pritchett, So Young Kim & Karen M Feigh, *Measuring Human-Automation Function Allocation*, 8 J. COGN. ENG. DECIS. MAK. 52–77 (2014). David D Woods, *Cognitive technologies: The design of joint human-machine cognitive systems*, 6 AI MAG. 86 (1985).

[212] Madeleine Clare Elish Elish, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, 5 ENGAG. SCI. TECHNOL. SOC. 40–60 (2019). (describing the moral crumple zone as occurring when the responsibility is misattributed to a human actor who had limited control over the behavior of the automated system).

"responsibility gap,"[213] but can be simply described as "blame the human." As James Reason explains:

> "The occurrence of a man-made disaster leads inevitably to a search for human culprits. Given the ease with which the contributing human failures can subsequently be identified, such scapegoats are not hard to find. But before we rush to judgement there are some important points to keep in mind. First, most of the people involved in serious accidents are neither stupid nor reckless though they may well have been blind to the consequences of their actions. Second, we must beware falling prey to the fundamental attribution error (i.e. blaming people and ignoring situational factors). As Perrow argued, it is in the nature of complex, tightly-coupled systems to suffer unforeseeable sociotechnical breakdowns. Third, before beholding the mote in his brother's eye, the retrospective observer should be aware of the beam of hindsight bias in his own."[214]

Whether one wants to mask the discrimination, blame the machine, or blame the human, the rest of this Essay will show how anti-discrimination law's failure to understand the interdependency and complexity of cybernetic systems creates a black hole which anyone can use to evade accountability.

IV. Disparate Treatment and Disparate Impact Liability Cannot Cope with Cybernetic System Discrimination

The interdependent and complex nature of cybernetic systems is incompatible with the way the law understands discrimination to occur. While there have been debates about which liability models are best suited to litigate human discrimination, machine discrimination, or systemic discrimination, this section shows that that debate is misguided. Disparate treatment (including the Equal Protection Clause's intentional discrimination) and disparate impact, the two key mechanisms of discrimination liability, do not even comprehend how discrimination occurs in cybernetic systems – making them a black hole. Without a meaningful path to accountability for discrimination, this will continue to incentivize the use of cybernetic systems as a way to insulate from liability.

To understand the principles of disparate treatment and disparate impact, this section focuses on Title VII of the Civil Rights Act of 1964.[215] Specifically, Title VII makes it unlawful for employers to use employment practices that have an adverse effect on members of a certain race, color, religion, sex, or national origin because of the employee's protected class (disparate

[213] Müller, *supra* note 206.

[214] Reason, *supra* note 205 at 216. (citing Perrow, *supra* note 199.

[215] 42 U.S. Code § 2000e–2 ("(a) Employer practices: It shall be an unlawful employment practice for an employer— (1) to fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin; or (2) to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin.")

treatment) or as compared to members of another race (disparate impact).[216] Employees and applicants are protected from discriminatory decisions including failing or refusing to hire, compensation, terms, conditions, privileges, and limiting, segregating or classifying employees or applicants in any way that would deprive any individual of employment opportunities or otherwise adversely affect their status.[217]

To show the cybernetic black holes within antidiscrimination law, this section identifies the four assumptions inherent to both disparate treatment and disparate impact law and then reveals how they conflict with the interdependence and complexity of cybernetic systems they are supposed to be able to regulate. As shown in Table 1, the sum of legal and technical barriers that plaintiffs must overcome to show disparate treatment or disparate impact transforms cybernetic systems into black holes for antidiscrimination analysis – a place to where discrimination can occur without the law ever recognizing it. This cybernetic reality means that disparate treatment and disparate impact cannot be leveraged by plaintiffs against cybernetic system discrimination leaving those discriminated against without remedy and those even thinking about avoiding discrimination without any incentive to do so. In other words, the anticlassification perspective sanctions discrimination in our cybernetic world, and moreover, incentivizes users and organizations to adopt even more cybernetic systems in order to shield themselves from liability.

[216] 42 U.S.C. § 2000e–2(k)(1)(A)(i). *See also, Albemarle Paper Co. v. Moody*, 422 U.S. 405 (1975); *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971).

[217] 42 U.S.C. § 2000e–2(a) ("It shall be an unlawful employment practice for an employer— (1) to fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin; or (2) to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin.")

Table 1. Summary of the assumptions inherent to disparate treatment and disparate impact and how they conflict with the reality of cybernetic system discrimination.

Disparate treatment Assumptions	**Cybernetic Reality**
Intentionally caused discrimination against someone based on their protected class	Behavior is often emergent, unpredictable, and drift towards failure without any actor able to exercise enough control to *intentionally cause* discrimination.
By rationally and invidiously considering their protected class	Finding the specific
At the moment of decision	There is no single moment of decision. Instead, there are numerous decisions by numerous actors all contributing their part to discrimination.
While in complete control of their decision-making process	The decision-making is made jointly by the human and machine and influenced by the organization such that none are in complete control.

Disparate Impact Assumptions	**Cybernetic Reality**
An unfair machine or human's test has discriminatory results	Even tests evaluated to show fairness prior to deployment cannot ensure fairness once deployed in the real world. Tests are administered through machines and humans, within organizations which can each undermine its validity once deployed.
Caused by a specific failure	Each actor contributes to the performance of the test, often in complex ways dependent on each other, such that there is no single, specific failure.
Identifiable prior to deployment	Behavior is often emergent, unpredictable, and drifts towards failure and even experts have yet to establish methods of comprehensively identifying sources of failure prior to deployment.
And the employer refuses to use an alternative employment practice that exists, is available, equally valid, and less discriminatory	This is asking an individual plaintiff to succeed where entire research fields have yet to succeed: overcome the barriers to cybernetic system performance.

A. Disparate Treatment

Disparate treatment occurs where there is evidence of a discriminatory motive.[218] Specifically under Title VII, it is unlawful to "discriminate against any individual with respect to… employment, *because of* such individual's race, color, religion, sex, or national origin[.]"[219] Unlike disparate impact, the employment practice in disparate treatment is not facially neutral. Broadly speaking, in defining disparate treatment, the Supreme Court and numerous Circuit Courts have held or suggested that prohibition against disparate treatment is the same as the Constitutional prohibition against intentional discrimination under the Fourteenth Amendment's Equal Protection Clause.[220] But whether under equal protection or disparate treatment, this theory has been extensively criticized as being incapable of addressing human discrimination, machine discrimination, or systemic discrimination.[221] Courts are "hostile to discrimination cases" establishing "Losers' Rules that serve to justify prodefendant outcomes,"[222] and that have been empirically shown to create nearly insurmountable barriers to plaintiff's claims.[223] This subsection will show that the assumptions underlying disparate treatment law cannot cope with cybernetic system discrimination either.

To prove disparate treatment, the plaintiff must first make a prima facie claim of discrimination by preponderance of the evidence.[224] To do this, the plaintiff must show they: (i) belong to a protected class, (ii) were qualified for whatever was denied to them, (iii) were subjected to an adverse employment action, and (iv) the employer gave better treatment to similarly situated person outside the plaintiff's class.[225]

Plaintiffs can allege disparate treatment as the product of either single or mixed motives. In single source cases, also known as pretextual discrimination cases, the focus is to examine a "single source"[226] to determine "whether either illegal or legal motives, but not both, were the 'true'

[218] *Int'l Bhd. of Teamsters v. United States* 431 U.S. 324, 335 n. 15 (1977). *See also*, Barocas and Selbst, *supra* note 56 at 696.

[219] 42 U.S.C. § 2000e–2(a)(1) (emphasis added).

[220] The initial landmark decision was *Washington v. Davis*, 426 U.S. 229 (1976). Following cases have read the two doctrines as virtually equivalent. *Int'l Bhd. of Teamsters v. United States* 431 U.S. 324, 335 (1977); *LeBlanc-Sternberg v. Fletcher*, 67 F.3d 412, 425–26 (2d Cir. 1995); *Grano v. Dep't of Dev. of City of Columbus*, 637 F.2d 1073, 1081 (6th Cir. 1980); *Gallagher v. Magner*, 619 F.3d 823, 831 (8th Cir. 2010); *Pac. Shores Properties, LLC v. City of Newport Beach*, 730 F.3d 1142, 1158 (9th Cir. 2013); *His House Recovery Residence, Inc. v. Cobb Cty., Georgia*, 806 F. App'x 780, 784 (11th Cir. 2020); *2922 Sherman Ave. Tenants' Ass'n v. D.C.*, 444 F.3d 673, 684 (D.C. Cir. 2006). See also, Richard A Primus, *The Future of Disparate Impact*, 108 MICH. LAW REV. 48, 1361–62 (2010). ("Even if the *Ricci* Court had kept scrupulously to the terminology of disparate treatment doctrine, the substance of its analysis would have been largely transferable to the equal protection context. That the Court did not even bother to keep the terminologies separate only testifies to the artificiality of the distinction between them in practice. So despite the Court's presentation of the *Ricci* premise as a matter of statutory law only, one can probably substitute "equal protection" for "disparate treatment" and have an equally valid proposition.")

[221] *Supra* Sec. I.

[222] Nancy Gertner, *Loser's Rules*, 122 YALE LAW J. FORUM 109, 109–10 (2012).

[223] Michael J. Zimmer, *The New Discrimination Law: Price Waterhouse is Dead, Whither McDonnell Douglas?*, 53 EMORY LAW J. 1887, 1943–44 (2005). (collecting empirical studies)

[224] *Burdine*, 450 U.S. at 253 (1981)

[225] *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 802 (1973).

[226] *Price Waterhouse v. Hopkins*, 490 U.S. 228, 247 (1989) (describing the basic "premise [in pretext cases like] *Burdine* is that *either* a legitimate *or* an illegitimate set of considerations led to the challenged decision.") (citing *Texas Dep't of Cmty. Affairs v. Burdine*, 450 U.S. 248 (1981))

motives behind the decision."[227] If the plaintiff makes the prima facie claim under this framework, originally laid down in *McDonnell Douglas Corporation v. Green*,[228] the burden then shifts to the employer to "articulate some legitimate, nondiscriminatory reason."[229] The employer need not have made "the *best* decision, it simply must [have made] a legitimate decision untainted by illegitimate motives."[230] If the employer articulates such a reason, then the burden shifts back to the plaintiff, "to show that [the] stated reason for the adverse employment decision] was in fact pretext."[231] If the plaintiff can show pretext, they will have proved a Title VII pretextual discrimination claim.

In mixed-motives cases, however, there is no one "true" motive behind the decision. So, when a plaintiff makes the prima facie case against "decisions based on a mixture of legitimate and illegitimate considerations" it will be analyzed on the *Price Waterhouse* framework.[232] Here, Title VII does not "obligate a plaintiff to identify the precise causal role played by legitimate and illegitimate motivations in the employment decision she challenges. [Instead, Title VII only] obligate[s] her to prove that the employer relied upon [illegitimate] considerations in coming to its decision."[233] To rebut this, the employer need only show under a preponderance of evidence that it would have "made the same decision even if it had not allowed [illegitimate considerations] to play such a role."[234]

In practice, courts have intermittently excluded or disregarded evidence of discriminatory intent without a coherent rationale – often "exaggerate[ing] the costs of allowing evidence to be considered or minimiz[ing] the benefits from doing so."[235] In addition to the uncertainty around what evidence certain courts will accept, former District Court Judge Nancy Gertner explains that the asymmetric rules and heuristics of summary judgment mean that plaintiffs must "bear the burden of proving all elements of the claim, particularly intent, and must do so based on undisputed facts" much "earlier on in the litigation process than before, with far, far less information."[236]

In her seminal work, civil rights lawyer and law professor, Linda Hamilton Krieger, outlined four critical assumptions that courts make in disparate treatment cases and then proceeded to show

[227] *NLRB v. Transportation Management Corp.*, 462 U.S. 393, 400, n. 5 (1983).

[228] *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 802 (1973)

[229] *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 802 (1973)

[230] Carla A. Ford, *Gender Discrimination and Hostile Work Environment*, 57 US ATTY. BULL. 1, 2 (2009). (citing *Pottenger v. Potlatch Corp.*, 329 F.3d 740, 748 (9th Cir. 2003) (employer has "leeway to make subjective business decisions, even bad ones.")

[231] *McDonnell Douglas Corp.*, 411 U.S. 792, at 804.

[232] *Price Waterhouse v. Hopkins*, 490 U.S. 228, 241 (1989)

[233] *Price Waterhouse v. Hopkins*, 490 U.S. 228, 241–42 (1989)

[234] *Price Waterhouse v. Hopkins*, 490 U.S. 228, 244–45 (1989) ("We think these principles require that, once a plaintiff in a Title VII case shows that gender played a motivating part in an employment decision, the defendant may avoid a finding of liability only by proving that it would have made the same decision even if it had not allowed gender to play such a role. This balance of burdens is the direct result of Title VII's balance of rights.") See also, *Mt. Healthy City Bd. of Ed. v. Doyle*, 429 U.S. 274, 287 (1977), applied in *Givhan v. Western Line Consolidated School Dist.*, 439 U.S. 410, 417 (1979); *Arlington Heights v. Metropolitan Housing Development Corp.*, 429 U.S. 252, 270–271, n. 21 (1977); *Hunter v. Underwood*, 471 U.S. 222, 228 (1985).

[235] Aziz Z Huq, *What is Discriminatory Intent?*, 103 CORNELL LAW REV. 83, 1267 (2020).

[236] Nancy Gertner, *supra* note 223 at 114–15. *See also* Elizabeth M. Schneider, *The Changing Shape Of Federal Civil Pretrial Practice: The Disparate Impact On Civil Rights And Employment Discrimination Cases*, 158 UNIV. PA. LAW REV. 517 (2010).

how those assumptions were invalid under cognitive psychology's understanding of cognitive bias.[237] Courts understand disparate treatment discrimination as occurring when a human (i) intentionally caused discrimination against someone based on their protected class (ii) by rationally and invidiously considering their protected class, (iii) at the moment of the decision, (iv) while in complete control of their decision-making process.[238] This subsection will take the same four critical assumptions and show that they are also invalid given the scientific and engineering understanding of interdependent and complex cybernetic systems.

1. Intentionally Caused Discrimination Against Someone Based on Their Protected Class

a. The Assumption

Section 703 of Title VII states that "[i]t shall be an unlawful employment practice for an employer… to fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin."[239] Although this language would seem to only require proof of causation and not intent, that is not how it has been construed. A plaintiff must do more than establish that their protected status "made a difference" or "played a role." [240] Under both "single motive" *McDonnell Douglas* cases and "mixed-motive" *Price-Waterhouse* and *Desert Palace* cases, the plaintiff can establish liability only if the one reason, or one of the reasons, respectively, for the adverse employment decision was purposeful and intentional discrimination.[241]

Justice Brennan's plurality opinion in *Price Waterhouse* shows how courts have equated the causation analysis ("playing a motivating part" in a decision) with intent analysis (the conscious use of the plaintiff's group status in their decision):

> "In saying that gender played a motivating part in an employment decision, we mean that, if we asked the employer at the moment of the decision what its reasons were and if we received a truthful response, one of those reasons would be that the applicant or employee was a woman."[242]
>
> "Remarks at work that are based on sex stereotypes do not inevitably prove that gender played a part in a particular employment decision. The plaintiff must show that the employer actually relied on her gender in making its decision."[243]

[237] Krieger, *supra* note 22.

[238] This list is similar to the summary by former Federal District Judge Nancy Gertner. Nancy Gertner, *supra* note 223 at 120 n. 46. ("Even remarks that are 'arguably probative of bias' may not be probative at all unless they were (a) related to the employment, (b) made close in time to the employment decision, (c) uttered by decisionmakers or those in a position to influence the decisionmaker, and (d) unambiguous.") (citations omitted)

[239] 42 USC 2000e-2(a)(1)

[240] Krieger, *supra* note 22 at 1168.

[241] Zimmer, *supra* note 224 at 1925. Krieger, *supra* note 22 at 1168 n. 16 (collecting cases).

[242] *Price Waterhouse* 490 U.S. at 240.

[243] *Price Waterhouse* 490 U.S. at 252.

Even in Justice O'Connor's influential concurrence, she emphasizes that the plaintiff's key burden is to show evidence of "discriminatory animus in the decisional process."[244] In sum, "[d]iscrimination, even when subtle and unconscious, is assumed to result from discriminatory motive or intent."[245] This assumption is made explicit in age discrimination cases under the Age Discrimination in Employment Act where plaintiffs must show that age was the "but-for" cause of the adverse action.[246]

b. The Reality

Proving intent or causation in a cybernetic system is virtually impossible, let alone both at the same time. First, there is likely no intention to discriminate or cause the discrimination, and secondly, even if there is an intention that motivated someone to cause discrimination, that would be nearly impossible to find.

The emergent behavior of cybernetic systems means that many failures cannot be predicted ahead of time and are, in fact, surprises.[247] Users are often faced with large problem spaces, and many systems will drift towards failure where the users are unaware the system is even failing.[248] In the user's words, they may have thought the discrimination was impossible. Wagenarr and Groeneweg introduced the term *impossible accident* to describe what they found after reviewing 100 shipping accidents:

> "Accidents appear to be the result of highly complex coincidences which could rarely be foreseen by the people involved. The unpredictability is caused by the large number of causes and by the spread of the information over the participants. … Accidents do not occur because people gamble and lose, they occur because people do not believe that the accident that is about to occur is at all possible."[249]

Charles Perrow famously went further to argue that accidents were *normal* in cybernetic systems,[250] that it is the "nature of complex, tightly-coupled systems to suffer unforeseeable sociotechnical breakdowns."[251] Both of the accounts of accidents as normal or impossible are far too quick to dismiss the reality that many do "gamble and lose" and that most accidents have a history where disaster, or discrimination, could have been mitigated.[252] However, it still is important evidence to show that even in a hypothetical colorblind dreamland, where no one has intent to discriminate, there will be failures, there will be discrimination.

[244] *Price Waterhouse* 490 U.S. at 278. *See also, Price Waterhouse* 490 U.S. at 276 (Justice O'Connor explaining that the key inference is about if the "employer's discriminatory animus made a difference to the outcome")

[245] However, Courts do allow unconscious bias to create liability under disparate treatment however, unsurprisingly, they seem only interested in doing so for age discrimination and somewhat for gender – excluding race and national origin from that privilege. Krieger, *supra* note 22 at 1166–67.

[246] *Gross v. FBL Fin. Servs., Inc.*, 557 U.S. 167, 177–78 (2009).

[247] *Supra* Sec. III.C. 1

[248] *Supra* Sec. III.C. 2.

[249] REASON, *supra* note 205 at 216. (quoting Willem A. Wagenaar & Jop Groeneweg, *Accidents at sea: Multiple causes and impossible consequences*, 27 INT. J. MAN-MACH. STUD. 587 (1987).)

[250] PERROW, *supra* note 199.

[251] REASON, *supra* note 205 at 216. (citing PERROW, *supra* note 199.)

[252] BARRY TURNER, MAN-MADE DISASTERS (1978); REASON, *supra* note 49.

But imagine the cases where there was someone who purposefully and intentionally caused discrimination to occur. In the study of accidents, there are four main categories of human action: (1) non-intentional (involuntary) action, (2) unintentional action where actions fail to go as intended (slips and lapses), (3) intentional but mistaken action where intended actions fail to achieve their desired consequences and (4) intentional successful action.[253] Flipping these four categories of intentional actions above to represent someone with the intention to discriminate, it shows that for an invidious actor to be liable under disparate treatment law, they must not only have the intent but the ability to understand how the cybernetic system operates to the degree that they can cause the desired outcome of discrimination to occur. Only the intentional successful action here is cognizable by the disparate treatment or intentional discrimination understanding of anti-discrimination. For the law to see the discrimination under disparate treatment, it cannot be involuntary, it cannot be unintentional, it cannot be intentional but mistaken. No, the only justiciable discrimination under disparate treatment is discrimination from intentional and successful actions. To require plaintiffs to prove this level of intent and success within emergent, drifting cybernetic systems after the fact is nearly impossible.

By virtue of the realities above, even if the best researchers and lawyers investigated the cybernetic system knowing there was discrimination somewhere to be found in the system, it seems impossible to think they would ever find the smoking gun. This is especially true considering the reality that defendants could point to the interdependent and complex nature of cybernetic systems where discrimination failures are impossible to predict or simply, normal events. Even if there was intentionally caused discrimination, the defendant’s claims align perfectly with the aspirations of our anti-discrimination law to live in a colorblind dreamland.

So, to determine if someone was purposefully and intentionally causing discrimination to occur, the plaintiffs claiming discrimination must (a) find the actor with the intent, (b) show that the actor knew how to realize that intent, and (c) took actions that caused discrimination. Just to find the actor with intent would require assessing the myriad amount of people often involved in cybernetic systems, from the inputs, machine, users, and organization level, all distributed across time and location.[254] Imagine if the plaintiffs knew the designers of the test in *Ricci*, IOS, had wanted to discriminate against Black and Hispanic firefighters. Given the numerous factors that likely contributed to the discriminatory outcomes, the plaintiffs would have to provide evidence showing that IOS personnel intentionally designed a facially neutral exam to disproportionately benefit those with access to those with the study materials, to ask questions not germane to New Haven, and to reward memorization. It would be even more difficult if multiple people within IOS collected the background information for the exams and contributed to designing the exam. Then the presumably lone actor would have to have accounted for all the other people who courts would likely argue would have prevented these flawed had they known it would cause discriminatory outcomes.

Furthermore, IOS is only the “designer” element in the cybernetic system. Their defense is to merely identify the other elements of the cybernetic system and create ambiguity by pointing to other major elements contributing to this discriminatory accident: New Haven could have helped

[253] REASON, *supra* note 205 at 6.
[254] *Supra* Part III.B. 2.c

with access to study materials (machine); New Haven reviewed the exam ahead of time and actually operated the exam (user); New Haven chose IOS to build the exam using only writing and oral assessments when other methods are known to be less discriminatory (feedback and organization). Then, imagine a firefighter is suing New Haven as the defendant in this case. New Haven's defense is to point to IOS as the designer who they trusted to provide a neutral exam. As Samuel Bagenstos summarized, "it may be difficult, if not impossible, for a court to go back and reconstruct the numerous biased evaluations and perceptions that ultimately resulted in an adverse employment decision."[255] In cybernetic system discrimination, almost every element can reasonably be to blame for contributing to the output.

2. By Rationally and Invidiously Considering Their Protected Class

a. The Assumption

Under the *McDonnell Douglas* framework, "virtually all individual treatment cases turn on the third step…: proof of pretext."[256] After the employer "articulate[s] some legitimate, nondiscriminatory reason",[257] the plaintiff can prevail only by proving that the reason offered was "phony," a "sham," a "mask," a "façade," or a "cover-up" for the employer's "true" discriminatory motive.[258] The assumption here is that if the employer does not harbor discriminatory intent, they will act objectively and rationally, producing decisions without discriminatory outcomes.[259] As illustrated by the law and economics "taste for discrimination" perspective, a discriminating employer could and would be objective but for their overwhelming desire to discriminate against a protected class.[260] The employer was about to reach an objective rational decision but at the moment of decision consciously relied upon beliefs about the employee's protected class. This assumption pushes the plaintiff and defendant into a zero-sum, winner-take-all game: the non-discriminatory reason that the employer will offer as justification for the outcome is either real or phony, true or false. "Within the pretext paradigm, it is simply not possible for an employment decision to be both motivated by the employer's articulated reasons and tainted by intergroup bias; the trier of fact must decide between the two."[261] Plaintiff's attorneys cannot argue that "employer was a well-intentioned good person who, through lack of care did a bad thing."[262] "In the stories told by disparate treatment caselaw, there is no discrimination without an invidiously motivated actor. Every successful disparate treatment story needs a villain."[263]

b. The Reality

Imagine the plaintiffs were able to show intentionally designed the test in that way and that those decisions caused the discriminatory outcome. Now the plaintiffs must show that the designers not only intentionally caused the system to operate in a discriminatory way, but that they

[255] Bagenstos, *supra* note 31 at 9.

[256] Krieger, *supra* note 22 at 1178.

[257] *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 802 (1973)

[258] Krieger, *supra* note 22 at 1178 (internal quotations and citations omitted).

[259] *Id.* at 1167.

[260] *Id.* at 1182 n. 83. *See also*, Baert Stijn & Ann-Sophie De Pauw, *Is Ethnic Discrimination Due to Distaste or Statistics?*, 125 ECON. LETT. 270 (2014).

[261] Krieger, *supra* note 22 at 1179 (citations omitted).

[262] *Id.* at 1180–81.

[263] *Id.* at 1166–67.

intentionally caused discrimination *by considering protected class*. Once again, this is nearly impossible to show because often there will not be intentional consideration of protected class information and because even if they did, plaintiffs would never be able to find it.

3. At the Moment of Decision

a. The assumption

In analyzing whether bias-revealing remarks can support a finding of discrimination, courts believe all the information relevant to the determination of disparate treatment is contained in the "moment of decision," the "'snapshot' of the decisionmaker's mental state at the moment the allegedly discriminatory decision was made."[264] Courts have made clear that the person making remarks must have participated in, or influenced, the decisionmaking process, and the remarks must relate in timing and subject-matter to that decision.[265] A majority of justices in *Price Waterhouse* agreed with this "moment of decision" framework, with the plurality explaining that

> "[t]he present, active tense of the operative verbs of § 703(a)(1) ("to fail or refuse [to hire or to discharge any individual, or otherwise to discriminate against any individual]") … turns our attention to the actual moment of the event in question, the adverse employment decision. The crucial inquiry, the one commanded by the words of § 703(a)(1), is whether [the protected class] was a factor in the employment decision at the moment it was made."[266]

Justice O'Connor's concurrence reiterated this focus on the "moment of decision," stating that "stray remarks in the workplace, … statements by nondecisionmakers, or statements by decisionmakers unrelated to the decisional process itself" are insufficient "to satisfy the plaintiff's burden" in mixed-motives analysis.[267] Said a different way, according to Justice O'Connor, the only statements that are sufficient to trigger mixed-motives analysis are those related to the decision process itself.

While language like "moment" and "snapshot" could be read as figurative speech indicating a brief period of time and scope, most courts have followed *Price Waterhouse*'s lead by interpreting the "moment of decision" in shockingly narrow ways.[268] In *Heim v. State of Utah*, where the plaintiff, Debbie Heim, alleged sex discrimination against her supervisor, Mr. Tischner, the Tenth Circuit accepted that Mr. Tischner "in an angry outburst in the context of alleged problems with Ms. Heim's work, … remark[ed]: 'Fucking women, I hate having fucking women in the office.'

[264] *Id.* at 1183 n. 85. (citing *Sabree v. United Bhd. of Carpenters & Joiners Local* 33, 921 F.2d 396, 403-04 (1st Cir. 1990) (stating that *Price Waterhouse* directs the trial court "to essentially take a snapshot at the moment of the allegedly discriminatory act")).

[265] Daniel L. Kresh, Annotation, *Identity of Commenter and Relationship of Remark to Employment Decision as Determinants of Relevance of Stray Remark or Comment in Title VII Action for Sex Discrimination*, 4 A.L.R. Fed. 3d Art. 7 (2015) (II. Allegedly Stray Comments by Decision-Makers Relating to Adverse Employment Decision § 4. Allegedly stray comments by decision-makers relating to adverse employment decision are evidence of discrimination.)

[266] *Price Waterhouse v. Hopkins*, 490 U.S. 228, 240-41 (1989)

[267] *Price Waterhouse v. Hopkins*, 490 U.S. 228, 240-41 (1989)

[268] Krieger, *supra* note 22 at 1183 n. 88 (collecting cases). (also 1184)

Shortly after this outburst, Ms. Heim was refused permission to undertake a temporary field assignment for which she had previously been granted permission."[269] However, in affirming the trial court's refusal to proceed on mixed motives, the Tenth Circuit stated that "although the remark… was certainly inappropriate and boorish," it was merely "a statement of Mr. Tischner's personal opinion. The evidence does not show Mr. Tischner acted with discriminatory intent, only that he unprofessionally offered his private negative view of women during a display of bad temper at work."[270] Heim had failed to show a nexus between Mr. Tischner's "private" bias against women and the employment decision "shortly after."[271]

Moreover, if the plaintiff cannot identify the specific time of bias-revealing remarks, courts often presume, "absent admissible evidence to the contrary, that the [bias-revealing remarks] and the event were separated by enough time to render the remark nonprobative of discrimination."[272] In *Ferrand v. Credit Lyonnais*, Maia Ferrand brought a sex discrimination case showing that her immediate superior, Mr. Whitehair, "frequently refer[red] to Ferrand as a 'bitch,' 'whore,' or 'slut'… " in conversations with his superior, Mr. Ladouceur.[273] The Second Circuit affirmed the district court's decision to disregard the remarks because Mr. Ladouceur did "not indicate with any specificity when these allegedly discriminatory remarks occurred… . Such testimony does not provide the temporal or causal connection necessary to sustain a finding of pretext in this case."[274]

b. The reality

There is no moment of decision. In Krieger's original work about human cognition, she showed that "[w]hen interpersonal judgment is understood as an integrated system involving perception, interpretation, attribution, memory, and decisionmaking, the distinction between stereotype-revealing comments made during decisionmaking and before decisionmaking utterly breaks down."[275] The cybernetic perspective here expands this integrated system to include the machine and the organization to show that a belief in a "moment of decision" is even more archaic. For example, imagine the firefighter exam in *Ricci* was the product of intentional discrimination. Where would the moment of decision be found?[276] Was the moment when the IOS designed a written test? They are widely known to be discriminatory, but it was New Haven's union contract from years earlier that required the high-weighting of the likely-discriminatory written test. Was the moment when New Haven chose to use IOS who had never designed a promotion exam before? Or when IOS decided to focus on memorization or when they designed it incompatible with ordered ranking or pass-fail thresholds? Or when New Haven prevented themselves from being

[269] *Heim v. State of Utah*, 8 F.3d 1541, 1546 (10th Cir. 1993)

[270] *Heim v. State of Utah*, 8 F.3d 1541, 1547 (10th Cir. 1993)

[271] *Heim v. State of Utah*, 8 F.3d 1541, 1547 (10th Cir. 1993)

[272] Daniel L. Kresh, Annotation, *Identity of Commenter and Relationship of Remark to Employment Decision as Determinants of Relevance of Stray Remark or Comment in Title VII Action for Sex Discrimination*, 4 A.L.R. Fed. 3d Art. 7 (2015) (describing, in Part 1 §3, the importance of introducing evidence of when statements were made).

[273] *Ferrand v. Credit Lyonnais*, 2003 WL 22251313 at *1 (S.D. N.Y. 2003), summarily aff'd, 110 Fed. Appx. 160 (2d Cir. 2004) (unpublished opinion)

[274] *Ferrand v. Credit Lyonnais*, 2003 WL 22251313 at *12 (S.D. N.Y. 2003), summarily aff'd, 110 Fed. Appx. 160 (2d Cir. 2004) (unpublished opinion) (citing with approval, *Geier v. Medtronic, Inc.*, 99 F.3d 238, 242 (7th Cir. 1996) ("To be probative of discrimination, isolated comments must be contemporaneous with the [decision in question] or causally related to the ... decision making process."))

[275] Krieger, *supra* note 22 at 1185.

[276] *Supra* Sec. **Error! Reference source not found.**

able to check the content of the exams prior to using them? There are numerous decisions by numerous actors all contributing their part to the discriminatory outcome. There is no single moment of decision.

4. While in Complete Control of Their Decision-Making Process.

a. The Assumption

Krieger explains that "disparate treatment jurisprudence–indeed the entire normative structure of anti-discrimination law–is based on an assumption that decisionmakers possess 'transparency of mind,' that they can accurately identify why they are about to make, or have already made, a particular decision. According to this view, if an employee's protected group status is playing a role in an employer's decision making process, the employer will be aware of that role, even if he is not honest (or careless) enough to admit it. Equipped with conscious self-awareness, well-intentioned employers become capable of complying with the law's proscriptive injunction not to discriminate. They will monitor their decision making processes and prevent prohibited factors from affecting their judgments."[277] In fact, the entire concept of the "moment of decision" above relies on this assumption, that decisionmakers "have ready access to the workings of their own inferential process. If they simply chose to be truthful, they could tell us whether an employee's race, ethnicity, or gender had influenced their decision."[278] As British philosopher, Gilbert Ryle, explains, this assumption requires that a "person has direct knowledge of the best imaginable kind of the workings of his own mind. Mental states and processes are (or are normally) conscious states and processes, and the consciousness which irradiates them can engender no illusions and leaves the door open for no doubts. A person's present thinkings, feelings and willings, his perceivings, rememberings and imaginings are intrinsically "phosphorescent"; their existence and their nature are inevitably betrayed to their owner."[279]

b. The Reality

Simply put, no one is in complete control of a cybernetic system. At minimum, there are three foundational sets of humans coming together that could be the one theoretically in control – the user, the machine (and its designer), and the organization they operate within. The designers and machines may seem to have control but they only affect the outcome by acting through a third party, the user. Even if the data is biased or the design is flawed, it operates through a human being and an organization, not on its own. It may be being used for the wrong purpose or be fed the wrong information by the user. Similarly, the machine is interdependent on the human. They are fundamentally a joint system. The human is informed and influenced by the machine. The human may rely too little or too much on the machine. Third, the organization controls by guiding the designers explicitly through design requirements or implicitly by purchasing a certain version of the technology. The organization controls both the human and the machine by determining which machine is used in which situation, what resources and training is provided to the human including those regarding resources and training. This adds up to the definition of interdependence, where each are dependent on the other to produce outcomes, including failures like discrimination.

[277] Krieger, *supra* note 22 at 1167.

[278] *Id.* at 1185.

[279] GILBERT RYLE, THE CONCEPT OF MIND 154 (1949).

B. Disparate Impact

Disparate impact liability was established to protect people from "practices that are facially neutral in their treatment of different groups but that in fact fall more harshly on one group than another and cannot be justified by business necessity."[280] Disparate impact is used to prevent discrimination in employment generally and on the basis of age, discrimination against persons with disabilities, and discrimination in housing.[281] The establishment of disparate impact law "has been universally hailed as the most important development in employment discrimination law"[282] and "scholars have offered numerous proposals to extend the disparate impact theory to cure all manner of social ills."[283] But despite this optimism among some, others argue that disparate impact is "complicated and confusing,"[284] that the development of disparate impact law has contributed to increased discrimination,[285] that disparate impact claims are empirically "more difficult to prove than standard intentional discrimination claims,"[286] that, in practice, the intent requirement with all its problems has not truly been left behind,[287] and that attempting to leave intent behind was not even the correct goal in the first place.[288]

To prove a Title VII disparate impact violation, a plaintiff must establish a prima facie case by "offer[ing] statistical evidence of a kind and degree sufficient to show that the practice in question has caused the exclusion of applicants for jobs or promotions because of their membership in a protected group."[289] A plaintiff may establish a prima facie case in a number of ways, including by arguing that the selection procedure shows disparate impact under the EEOC Uniform Guidelines on Employee Selection Procedures.[290] Under the Uniform Guidelines, if one group's selection rate is less than 80 percent "of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact" – though higher rates could still be evidence of adverse impact.[291] Ultimately, the " 'significance' or 'substantiality'

[280] *Int'l Bhd. of Teamsters v. United States* 431 U.S. 324, 335 n. 15 (1977).

[281] Title VII of the Civil Rights Act of 1964 (Title VII), 42 U.S.C. § 2000e *et seq.*; the Americans with Disabilities Act of 1990 (ADA), 42 U.S.C. §§ 12112 (b)(2), (b)(6); the Age Discrimination in Employment Act (ADEA) 29 U.S.C. §621 *et seq.*; and, the Fair Housing Act (FHA), 42 U.S.C. §§ 804(a) and 805(a).

[282] Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, UCLA LAW REV. 83, 703 (2006).

[283] *Id.* at 704.

[284] CharlesA Sullivan, *The World Turned Upside Down: Disparate Impact Claims by White Males*, 98 NORTHWEST. UNIV. LAW REV. 1505, 1521 (2004).

[285] Harris and West-Faulcon, *supra* note 58.

[286] Selmi, *supra* note 283 at 734.

[287] Stacy E Seicshnaydre, *Is the Road to Disparate Impact Paved with Good Intentions: Stuck on State of Mind in Antidiscrimination Law*, 42 WAKE FOR. LAW REV. 1141 (2007).

[288] Selmi, *supra* note 283.

[289] *Watson v. Forth Worth Bank & Trust*, 487 U.S. 977, 994 (1988).

[290] *Watson v. Forth Worth Bank & Trust*, 487 U.S. at 995–996, n. 3 (plurality opinion) (EEOC's 80–percent standard is "a rule of thumb for the courts").

[291] 29 C.F.R. §1607.4(D) ("A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group.")

of numerical disparities [is judged] on a case-by-case basis."[292]

Once the plaintiff has established the prima facie case, the employer may rebut the prima facie case by "demonstrat[ing] that the challenged practice is *job related* for the position in question and *consistent with business necessity*[.]"[293] If the employer meets that burden, the plaintiff may still show a Title VII violation by demonstrating that the employer refuses to adopt an alternative employment practice that exists, is available, equally valid, and less discriminatory.[294]

As above for disparate treatment, this subsection outlines four critical assumptions that courts make in disparate impact cases and shows how they are invalid given our understanding of interdependent and complex cybernetic systems. In sum, courts understand disparate impact discrimination as occurring when (i) an unfair machine or human's test has discriminatory results (ii) caused by a specific failure, (iii) identifiable prior to deployment, (iv) and that there is an equally effective, less discriminatory alternative employment practice which the employer refused to adopt.

1. An Unfair Machine or Human's Test has Discriminatory Results

a. The Assumption

Disparate impact litigation, particularly in Title VII, has always centered on tests. In essentially every Supreme Court case ruling on Title VII disparate impact, there is an articulatable test at the center of the litigation: a written or oral test,[295] a requirement or condition on employment,[296] or a supervisor rating system.[297] Notably, successful challenges tended not to focus on the underlying test itself but instead the cut-off scores for acceptable or unacceptable employees.[298] In Title VII, Congress explicitly allowed employment tests but only if they were a "professionally developed ability test [and] not designed, intended or used to discriminate because of race, color, religion,

[292] *Watson*, 487 U.S. at 995 n.3 (citations omitted).

[293] 42 U.S.C. § 2000e-2(k)(1)(A)(i). The "touchstone" for disparate-impact liability is the lack of "business necessity": "If an employment practice which operates to exclude [minorities] cannot be shown to be related to job performance, the practice is prohibited." *Ricci*, 557 U.S. at 578 (quoting *Griggs*, 401 U.S. at 431). See also, *Griggs* 401 U.S. at 432 (stating that it is the employer's burden to demonstrate that practice has "a manifest relationship to the employment in question"); *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 425.

[294] 42 U.S.C. §§ 2000e–2(k)(1)(A)(ii) and (C); *Albemarle Paper Co. v. Moody*, 422 U.S. at 425. (allowing plaintiffs to show "that other tests or selection devices, without a similarly undesirable racial effect, would also serve the employer's legitimate interest")

[295] *Connecticut v. Teal*, 457 U.S. 440 (1982) (written examination); *Albemarle Paper Co. v. Moody*, 422 U.S. 405 (1975) (written aptitude tests); *Ricci v. DeStefano*, 557 U.S. 557 (2009) (written and oral examination); *Lewis v. City of Chicago, Ill.*, 560 U.S. 205 (2010) (the city's choice of cutoff for firefighter exam scores); *Griggs v. Duke Power Co.*, 401 U.S. 424, 431 (1971) (high school completion requirement).

[296] *Dothard v. Rawlinson*, 433 U.S. 321 (1977) (height and weight requirements); *New York City Transit Authority v. Beazer*, 440 U.S. 568 (1979) (rule against employing drug addicts); *Young v. United Parcel Serv., Inc.*, 575 U.S. 206 (2015) (weight lifting requirements for package deliverers); *Int'l Bhd. of Teamsters v. United States*, 431 U.S. 324 (1977) (seniority system and which jobs were available to who); *Griggs v. Duke Power Co.*, 401 U.S. 424, 431 (1971) (general intelligence test).

[297] *Meacham v. Knolls Atomic Power Lab'y*, 554 U.S. 84, 88 (2008) (supervisor rating employees on subjective measures of "performance," "flexibility," and "critical skills").

[298] Selmi, *supra* note 283 at 763 n. 225. ("successful challenges to tests tended to involve validating cut-off scores, as opposed to the underlying test itself").

sex or national origin."[299] This is the source of the flawed assumption: there is a test (or examination process) could have been sufficiently validated prior to deployment to know how it will perform in practice.

The Equal Employment Opportunity Commission (EEOC), established by the same Civil Rights Act of 1964 to "prevent any person from engaging in any unlawful employment practice" under Title VII,[300] has a similar singular focus on tests. In 1978, the EEOC adopted the Uniform Guidelines on Employee Selection Procedures or "UGESP."[301] As explained by the EEOC Office of Legal Counsel in 2006, "UGESP was published at a time when lawyers and psychologists were confronting the differences between judicial and scientific approaches to assessing the effects of employment tests. UGESP provided uniform federal government guidelines for establishing when employment tests were not discriminatory."[302] So although the regulations focus on "tests and other selection procedures,"[303] the definition of "selection procedure" is actually just a broader definition of tests.[304]

The main focus of the Guidelines is to assist employers in establishing criterion-related validity – that the tested elements of job performance are critical to the job and that the test effectively measures them.[305] To establish validity, industrial psychologists or other trained professionals perform a job analysis to identify the critical elements of job performance (job duties, work behaviors, and work outcomes) and then select and develop "measurable criteria that serve as metrics of how well an individual can perform the key functions of the job."[306] It is all about validating tests prior to deployment.

b. The Reality

The Court's and the EEOC's focus on tests alone and the belief that they can be fully validated prior to deployment is incredibly out of date.[307] As shown by *Ricci*, obsession with tests can make courts blind the cybernetic reality of how tests operate in the real world. The *Ricci* Majority even acknowledged this very problem when they quoted the foundational Title VII case of *Griggs v. Duke Power* where "the Court interpreted [Title VII] to prohibit, in some cases, employers' facially

[299] 42 U.S. Code § 2000e–2(h)

[300] 42 U.S. Code § 2000e–5(a)

[301] 29 CFR Part 1607.

[302] https://www.eeoc.gov/meetings/meeting-may-16-2007-employment-testing-and-screening/miaskoff

[303] 29 C.F.R. § 1607.1(A)

[304] 29 CFR § 1607.16(Q). ("Any measure, combination of measures, or procedure used as a basis for any employment decision. Selection procedures include the full range of assessment techniques from traditional paper and pencil tests, performance tests, training programs, or probationary periods and physical, educational, and work experience requirements through informal or casual interviews and unscored application forms.")

[305] *See generally* 29 C.F.R. § 1607.14 (2019) ("The following minimum standards, as applicable, should be met in conducting a validity study.").

[306] Matthew U Scherer, Allan G King & Marko N Mrkonich, *Applying Old Rules to New Tools: Employment Discrimination Law in the Age of Algorithms*, 71 S. C. LAW REV. 449, 479 (2019).

[307] Selmi, *supra* note 283 at 705. ("Outside of the original context in which the theory arose, namely written employment tests, the disparate impact theory has produced no substantial social change and there is no reason to think that extending the theory to other contexts would have produced meaningful reform. … Even with written tests the theory did not achieve the expected reform, as the vast majority of tests continue to have significant adverse impact.") (internal citations omitted)

neutral practices that, in fact, are 'discriminatory in operation.'"[308] In short, the Supreme Court in *Griggs* acknowledged that neutral practices could produce discriminatory outputs. But in the same breath, the *Ricci* Majority rejected this precedence. They claimed that New Haven "thought about promotion qualifications and relevant experience in neutral ways. They were careful to ensure broad racial participation in the design of the test itself and its administration. As we have discussed at length, the process was open and fair."[309] Because the process was designed to be open and fair, there could be no discriminatory impact.

Justice Ginsburg's dissent rightly accuses the Majority of "rest[ing] on the false premise that respondents showed "a significant statistical disparity," but "nothing more.""[310] Because, despite their references to *Griggs*, when it came to looking at the operation or the administration of the exam, the Majority's decision had already been made up once they deemed that the "process [for developing the exam] was open and fair." As described above, there were numerous issues with the operation and administration of the exam.[311] That the Court could not see the threats to the validity of the test once the test was "neutrally designed" and thus could not fit the discriminatory results of the test as cognizable under disparate impact. The Majority refused to see the possibilities that *Griggs* explicitly identified.

This reality is not surprising. As Scherer and his colleagues explain: "the Guidelines' forty-year-old standards are overdue for revamping or replacement to bring them in line with the modern social science of test validity, which has evolved considerably in the decades since the Guidelines first appeared. … the Guidelines and the existing case law on validation are bereft of meaningful discussion of these threats to validity."[312]

But the key threats to test validity are not simply questions of social science but of cognitive and systems engineering as well. The essence of cybernetic systems is that machines (tests) and their users exist within a complex set of interdependent relationships that adapt to and interact with the environment. Said differently, a neutrally designed test shown to have validity in theory, does not mean much until it is deployed in the real world. Until all the elements of the cybernetic system are put together and operated in the real world, it is hard to know exactly how they will perform. And even if one knows how the test *should* perform, that is very different than understanding how the test within the complete cybernetic system *performed* on a given day in a given environment. Just as "no plan survives first contact with the enemy,"[313] no theoretical test survives first contact with reality.

This issue is understood as the envisioned world problem.[314] In application to employment tests, the envisioned world problem would suggest that even a neutrally designed and valid test will face numerous challenges when actually deployed. For example, the test may be underspecified in that it is vague on many aspects of what it would be to administer the test in

[308] *Ricci*, 557 U.S. at 577–78 (quoting *Griggs v. Duke Power Co.*, 401 U.S. 424, 431 (1971))
[309] *Ricci v. DeStefano*, 557 U.S. 557, 592–93
[310] Ricci, 557 U.S. at 644
[311] Supra Sec. III.B.
[312] Scherer, King, and Mrkonich, *supra* note 307 at 482–83.
[313] https://hbr.org/2016/06/strategic-plans-are-less-important-than-strategic-planning
[314] Where the envisioned world describes the future technological systems (e.g., tests) in work domains that do not exist yet. Miller and Feigh, *supra* note 47 at 2–3. (internal citations omitted)

practice. In *Ricci*, there was unequal access to the study materials due to cost and delivery delays that could have affected the results.[315] Separately, the users or organization may be miscalibrated or overconfident that, if the test is merely administered as realized, the predicted consequences and only the predicted consequences will occur. Again, in *Ricci*, New Haven (and the Majority) seemed overconfident that a neutrally designed and valid written and oral exam would not have disparate outcomes, despite numerous professionals explaining that historically written exams produced disparities between races.[316]

Social scientists have been attempting to address these envisioned world challenges for many years despite the Guidelines, yet social science cannot fully identify the breadth of challenges cybernetic systems create. In 1999, the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education released the second edition of their "Standards for Educational and Psychological Testing" which is relied upon by many as a supplement to the Guidelines. Their standards discuss the importance of the user studying and evaluating the developer's materials,[317] following the developer's instructions during administration,[318] and only assuming or delegating responsibility to those "who have the training, professional credentials, and experience necessary to handle this responsibility."[319]

Beyond *Ricci*, there is a more infamous example in the employment testing world where a supposedly validated neutral test resulted in discriminatory outcomes: the General Aptitude Test Battery (GATB). The GTAB was developed in the 1940's by the U.S. Department of Labor for vocational counselling and job referral. In the 1980's the GATB was proposed as the single employment test for the U.S. Employment Service (USES) to screen approximately 19 million people for private- and public-sector jobs.[320] Only then, when the National Academy of Sciences (NAS) was tasked with reviewing its validity and potential for violating Title VII, were serious human-machine interaction and feedback problems that undermined the GATB's validity.

First, there were serious human-machine interaction and feedback problems. The USES trained their proctors to give honest answers when responding to questions regarding the scoring of questions.[321] This may not seem like an issue for most tests but the GATB (1) was a "speeded" two-choice (pick between two answers) test where no one was expected to finish if they spent time truly analyzing each question, and (2) had no penalty for incorrect answers, guessing or otherwise. As a result, the NAS showed that if an applicant just marked all of the answers the same way, they would get 50% correct and score higher than the 98th percentile.[322] Any test taker who asked about the scoring method almost assured themselves a place in the 98th percentile.

[315] Ricci 557 U.S. at 613-14, 17

[316] Ricci 557 U.S. at 572, 611-12.

[317] STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING, 111 (1999).

[318] *Id.* at 61. ("The usefulness and interpretability of test scores require that a test be administered and scored according to the developer's instructions. When directions to examinees, testing conditions, and scoring procedures follow the same detailed procedures, the test is said to be standardized. Without such standardization, the accuracy and comparability of score interpretations would be reduced.")

[319] *Id.* at 111.

[320] FAIRNESS IN EMPLOYMENT TESTING: VALIDITY GENERALIZATION, MINORITY ISSUES, AND THE GENERAL APTITUDE TEST BATTERY, 1338 vii (1989), http://www.nap.edu/catalog/1338 (last visited Jan 11, 2021).

[321] *Id.* at 100.

[322] *Id.* at 100.

Second, on a separate "speeded" part of the test, NAS found significant racial differences where white applicants finished more questions and earned higher scores than black applicants.[323] For the test, applicants had to find pairs of identical lines on a sheet and then search through a long list of answers to fill in the corresponding bubble. As a result, the test was not measuring an applicants ability to process information quickly, instead the test was measuring the amount of previous experience with tests, which correlated with different racial or ethnic groups.[324]

These issues were so problematic that the NAS called for a "vigorous program of research and development" and that these "[t]wo inadequacies in the testing program must be corrected."[325] This is for a test that, at the time, had been deployed for countless applicants in the U.S. for approximately 40 years. A supposedly validated neutral test, despite its serious flaws undermining its validity and suggesting likely discriminatory outcomes, was used for decades by the federal government and others.

2. Caused by a Specific Failure

a. The Assumption

Once plaintiffs have identified the test, they must identify the cause of the failure. Again, and again, the Supreme Court has required a clear, identifiable cause of the failure. In *Smith v. City of Jackson, Mississippi*, the court rejected the plaintiff's discriminatory impact claim because they had "not identified any specific test, requirement, or practice within the pay plan that has an adverse impact on older workers."[326] In *Wal-Mart Stores, Inc. v. Dukes*, the Court declined to certify a class action lawsuit for gender and pay discrimination in part because "merely proving that the discretionary system has produced a racial or sexual disparity is not enough. The plaintiff must begin by identifying the specific employment practice that is challenged."[327] In *Hazelwood School District v. United States*, despite the statistical data showing "substantial" differences in the hiring of Black teachers,[328] the Court still required "further evaluation by the trial court" as to whether Hazelwood had "engaged in a pattern or practice of employment discrimination."[329] These further evaluations the Court demands are all in search of identifying the source of failure.[330]

[323] *Id.* at 106.

[324] *Id.* at 106.

[325] *Id.* at 282.

[326] *Smith v. City of Jackson, Miss.*, 544 U.S. 228, 241 (2005)

[327] *Wal-Mart Stores, Inc. v. Dukes*, 564 U.S. 338, 357 (2011) (quoting *Watson* at 994 and citing *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 656 (1989) approving the statement) (internal quotations removed)

[328] *Hazelwood Sch. Dist. v. United States*, 433 U.S. 299, 308-10 (1977) (discussing how Hazelwood's percentage of Black teachers was 1.4% in 1972-1973, 1.8% in 1973-74 when "the percentage of qualified [Black] teachers in the area was… at least 5.7%" and an adjacent labor market had 15.4% Black teachers.)

[329] *Hazelwood Sch. Dist. v. United States*, 433 U.S. 299, 313 (1977). Hazelwood Sch. Dist. 433 U.S. at 312 ("[S]tatistics… come in infinite variety. … [T]heir usefulness depends on all of the surrounding facts and circumstances.") (quoting *Int'l Bhd. of Teamsters v. United States*, 431 U.S. 324, 340).

[330] *Hazelwood Sch. Dist. v. United States*, 433 U.S. 299, 312 (1977) ("to what extent those policies have changed the racial composition of that district's teaching staff from what it would otherwise have been; to what extent St. Louis' recruitment policies have diverted to the city, teachers who might otherwise have applied to Hazelwood; [and,] to what extent [Black] teachers employed by the city would prefer employment in other districts such as Hazelwood.")

In *Watson*, the Supreme Court plurality articulated a "causation"[331] requirement in disparate impact cases—affirmed later by *Wards Cove*[332] and *Inclusive Communities Project*[333]—stating that "the plaintiff's burden in establishing a prima facie case goes beyond the need to show that there are statistical disparities in the employer's work force. The plaintiff must begin by identifying the specific employment practice that is challenged. … Especially in cases where an employer combines subjective criteria with the use of more rigid standardized rules or tests, the plaintiff is in our view responsible for isolating and identifying the specific employment practices that are allegedly responsible for any observed statistical disparities."[334] The Court affirmed this requirement in *Wards Cove*, explaining that,

> "Our disparate-impact cases have always focused on the impact of *particular* hiring practices on employment opportunities for minorities. … As a general matter, a plaintiff must demonstrate that it is the application of a specific or particular employment practice that has created the disparate impact under attack. Such a showing is an integral part of the plaintiff's prima facie case in a disparate-impact suit under Title VII.
>
> [E]ven if on remand respondents can show that nonwhites are underrepresented… this alone will *not* suffice to make out a prima facie case of disparate impact. Respondents will also have to demonstrate that the disparity they complain of is the result of one or more of the employment practices that they are attacking here, specifically showing that each challenged practice has a significantly disparate impact on employment opportunities for whites and nonwhites. To hold otherwise would result in employers being potentially liable for the myriad of innocent causes that may lead to statistical imbalances in the composition of their work forces."[335]

The Court in *Inclusive Communities* continued that causation may be a matter of Constitutional significance: "[W]ithout [these] adequate safeguards [from requiring causation] at the prima facie stage, disparate-impact liability might cause race to be used and considered in a pervasive way and would almost inexorably lead governmental or private entities to use numerical quotas, and serious constitutional questions then could arise."[336]

[331] *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 656 (1989) (quoting affirmingly *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 994)

[332] *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 656 (1989) (quoting affirmingly *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 994)

[333] *Texas Dep't of Hous. & Cmty. Affs. v. Inclusive Communities Project, Inc.*, 576 U.S. 519, 542 (2015) ("A robust causality requirement ensures that "[r]acial imbalance ... does not, without more, establish a prima facie case of disparate impact" and thus protects defendants from being held liable for racial disparities they did not create.") (quoting *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 653 (1989), superseded by statute on other grounds, 42 U.S.C. § 2000e–2(k).)

[334] *Watson v. Fort Worth Bank & Tr.*, 487 U.S. 977, 994 (1988) (citing *Connecticut v. Teal*, 457 U.S. 440 (1982) for support).

[335] *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 656–57 1989. *Wards Cove Packing Co.*, 490 U.S. at 658 ("Consequently, on remand, the courts below are instructed to require, as part of respondents' prima facie case, a demonstration that specific elements of the petitioners' hiring process have a significantly disparate impact on nonwhites.")

[336] *Texas Dep't of Hous. & Cmty. Affs. v. Inclusive Communities Project, Inc.*, 576 U.S. 519, 542 (2015) (internal citations and quotations omitted). *See*, Reva B Siegel, *The Constitutionalization of Disparate Impact—Court-Centered and Popular Pathways: A Comment on Owen Fiss's Brennan Lecture*, 106 CALIF. LAW REV. 2001, 2014 (2018).

The Court in *Wards Cove* seemed to acknowledge that "[s]ome will complain that this specific causation requirement is unduly burdensome on Title VII plaintiffs" but then confidently stated that "liberal civil discovery rules" and the Guidelines requirements for employers to maintain records is sufficient to allow plaintiffs to "meet their burden of showing a causal link between challenged employment practices and racial imbalances in the work force."[337] Apparently the Court had never heard of the GATB.

Ultimately, if the plaintiffs cannot identify a specific cause of discrimination in the employment practice, then the Court will assume "no inference of discriminatory conduct."[338] The Court believes that "left to their own devices most managers in any corporation—and surely most managers in a corporation that forbids sex discrimination—would select sex-neutral, performance-based criteria for hiring and promotion that produce no actionable disparity at all."[339] And in an anti-classification framework whose goal is colorblindness, that belief is enough.

b. The Reality

It is nearly impossible to identify the specific cause of failure of a system. This is again, the myth of deconstruction believing that with enough access and discovery, enough questions, enough witnesses, the plaintiffs could determine the specific cause. This could not be further from the truth. Two of the most sophisticated investigations in the past 30 years, TWA 800 and Air France 447, never found the specific cause that caused the high-profile catastrophes with hundreds of casualties.

3. Identifiable Prior to Deployment

a. The Assumption

The Supreme Court requires that employers voluntarily comply with prohibitions on discrimination but prohibits them from making changes to the system after it has been deployed, evidencing the basic assumption that flaws in cybernetic systems can be identified prior to deployment. A central belief of the Court is that "Congress's intent [is] that voluntary compliance be the preferred means of achieving the objectives of Title VII. … [E]mployers' voluntary compliance efforts… are essential to the statutory scheme and to Congress's efforts to eradicate workplace discrimination."[340] In *Ricci*, the Court made clear that it is hesitant to hold employers liable for discriminatory impact when they could not have voluntarily complied *prior* to deploying

(discussing the implications of the causation requirement having constitutional significance)

[337] *Wards Cove Packing Co. v. Atonio,* 490 U.S. 642, 657–58 (1989)

[338] *Wal-Mart Stores, Inc. v. Dukes,* 564 U.S. 338, 355 (2011) (quoting *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 990 (1988))

[339] *Wal-Mart Stores, Inc. v. Dukes*, 564 U.S. 338, 355 (2011)

[340] *Ricci v. DeStefano,* 557 U.S. 557, 581, 583 (2009) (quoting *Firefighters v. Cleveland*, 478 U.S. 501, 515 (1986) and citing *Wygant v. Jackson Bd. of Ed.*, 476 U.S. 267, 290 (1986) (O'Connor, J., concurring in part and concurring in judgment)) (quotations omitted).

the test. The Court held that employers discovering and attempting to mitigate disparate impact results by altering the results or altering the test and re-running the applicants could be subject to disparate treatment liability.[341] In other words, "[m]aking changes after a tool has already been deployed is problematic under Ricci, which held that such modifications may be made only prospectively."[342] This is the ominous *Ricci* catch-22[343] where inaction leading to discrimination could lead to disparate impact liability and action to remedy that discrimination could lead to disparate treatment liability. The Court was much more concerned about the intentional reverse discrimination post-deployment reducing the number of white firefighters getting promoted, so it required employers to "have a strong basis in evidence to believe it will be subject to disparate impact liability if it fails to take the race-conscious discriminatory action."[344] Practically, given the difficulty of proving disparate impact, this requirement strongly incentivizes inaction on the part of employers. In response, the Court suggested pre-deployment testing as a way for employers to mitigate anticipated disparate impacts without necessarily violating Title VII.[345] Reenforcing their belief that discriminatory impacts can be prevented as long as there was enough pre-deployment testing.

b. The Reality

So, in addition to the challenges of identifying a specific cause of failure assuming that a single cause exists at all, courts now require that failures are identifiable prior to deployment – which is no easier. The reality of the envisioned world problem rears its head here, too. First, cybernetic systems have emergent drifting behavior that cannot always be predicted ahead of time. Scherer describes this emergence within the context of employment testing: "[I]t is almost inevitable that at least some disparate impacts will arise. Even if an employer succeeds in designing an algorithmic selection procedure that has no disparate impacts during initial training, adverse impacts may creep in as the characteristics of candidates and successful employees in a given position change."[346] This creeping behavior can be understood as the nature of cybernetic systems to drift towards failure, where the precursors of failure or discrimination is hard to identify prior to actual failure. Secondly, even expert scientists and engineers do not yet have the confidence that sources of failure for cybernetic systems can be reliably identified prior to deployment.[347] However, as usual, the Court has no issue demanding the near impossible from plaintiffs in discrimination cases.

[341] *Ricci* 557 U.S. at 629 (Alito, J., concurring).

[342] Scherer, King, and Mrkonich, *supra* note 307 at 496.

[343] *Id.* at 475.

[344] *Ricci v. DeStefano*, 557 U.S. 557, 585 (2009)

[345] Ricci 557 U.S. at 585 ("Title VII does not prohibit an employer from considering, before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race. And when, during the test-design stage, an employer invites comments to ensure the test is fair, that process can provide a common ground for open discussions toward that end. We hold only that, under Title VII, before an employer can engage in intentional discrimination for the asserted purpose of avoiding or remedying an unintentional disparate impact, the employer must have a strong basis in evidence to believe it will be subject to disparate-impact liability if it fails to take the race-conscious, discriminatory action.").

[346] Scherer, King, and Mrkonich, *supra* note 307 at 496.

[347] Infra Sec. V.A.

4. And That the Employer Refuses to Use an Alternative Employment Practice that Exists, is Available, Equally Valid, and Less Discriminatory

a. The Assumption

If the employer proves business necessity, as they often do,[348] the plaintiff can only prevail by showing that the employer has refused to adopt an alternative employment practice which would satisfy the employer's legitimate interests without having a disparate impact on a protected class, known as less discriminatory alternatives.[349] After *Griggs*, federal courts of appeals had initially required the defendant to show the absence of less discriminatory alternatives.[350] However, the Supreme Court quickly began holding[351] and the Civil Rights Act of 1991 codified[352] that this is the plaintiff's burden. In addition, the plaintiff must do more than merely allege such an alternative. The plaintiff must demonstrate that the alternative is equally effective and less discriminatory.[353] Moreover, the Supreme Court in *Wards Cove* determined that "cost or other burdens of the [plaintiff's] proposed selection devices are relevant in determining whether they would be equally as effective as the challenged practice in serving the employer's legitimate business goals."[354] Ultimately, the Court advises that the judiciary "are generally less competent than employers to restructure business practices; consequently, the judiciary should proceed with care before mandating that an employer must adopt a plaintiff's alternative selection or hiring practice in response to a Title VII suit."[355]

b. The Reality

In a cybernetic system, identifying the specific cause of the discriminatory impact is already an incredibly difficult burden. However, even if the specific cause is found, proposing and demonstrating a solution that would be equally difficult. The Court here is not only requiring the plaintiff to do the defendant's job but a job of expert scientists and engineers.

First, there is no way that a typical plaintiff can reasonably be expected to understand the complexities and interdependence of cybernetic systems in order to identify such a solution. In the

[348] Selmi, *supra* note 283 at 763 n. 225. (collecting cases to show that "there have been remarkably fewer testing cases in the last fifteen years, and courts increasingly have accepted employer justifications for their practices.")

[349] 42 U.S.C. § 2000e-2(k)(1)(A)(ii). *Albemarle*, 422 U.S. at 425 (requiring "other tests or selection devices, without a similarly undesirable racial effect")

[350] *Fitzpatrick v. City of Atlanta*, 2 F.3d 1112 (11th Cir.1993);

[351] *Albemarle Paper Co. v. Moody*, 422 U.S. 405 (1975); *Watson v. Ft. Worth Bank and Trust*, 487 U.S. 977 (1988); *Wards Cove Packing v. Atonio*, 490 U.S. 642 (1989).

[352] 42 U.S.C.A. § 2000e-2(k)(1)(C).

[353] *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 660 (1989) (in overcoming the business necessity defense, "respondents will have to persuade the factfinder that "other tests or selection devices, without a similarly undesirable racial effect, would also serve the employer's legitimate hiring interests; by so demonstrating, respondents would prove that petitioners were using their tests merely as a 'pretext' for discrimination.") (quoting *Albemarle Paper Co.*, supra, 422 U.S., at 425; and citing *Watson*, 487 U.S., at 998 (O'Connor, J.); id., at 1005–1006 (Blackmun, J., concurring in part and concurring in judgment).

[354] *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 661 (1989) (quoting *Watson*, supra, at 998 (O'Connor, J.).)

[355] *Wards Cove*, 490 U.S. 642, 661 (1989) (quoting *Furnco Construction Corp. v. Waters*, 438 U.S. 567, 578 (1978))

context of employment testing, identifying an equally valid and less discriminatory method of testing is the goal of the entire research community for decades – and yet, a typical plaintiff is supposed to be able to not only identify the solution but demonstrate that it is equally effective and less discriminatory. With the GATB, it took the National Academies of Science, with dozens of professional experts synthesizing hundreds of articles and new research studies over years to make recommendations of how to fix the test. And ultimately, the recommended a full redesign.[356] Apparently, every plaintiff should be able to do just the same.

Said differently, imagine that a plaintiff statistically showed that a specific medicinal drug was unsafe for a particular demographic. Using the disparate impact test, to get a remedy, the plaintiff would have to show not only that there was an alternative drug that existed, but that it was available, equally effective, and safer. Instead of merely requiring the plaintiff to show that the defendant inappropriately harmed them, the court is asking the plaintiff to take the job of a pharmaceutical company and medical doctors.

Separately, new researchers are starting to show that because of how race, gender and other types of discrimination are pervasive through our society, it is not clear the validity-diversity tradeoff cannot be overcome.[357] In other words, in many testing situations, the best predictors of performance are general intelligence which correlates with membership in protected classes.

V. SEARCHING FOR SOLUTIONS

This Essay has shown that our society is constituted by cybernetic systems where humans and machines make decisions within organizations and that the reality of how cybernetic systems operate and fail are fundamentally incompatible with our current antidiscrimination laws. The question then how to resolve this incompatibility? The following Subsection V.A. surveys the current recommendations spanning technical and legal solutions around fairness, access, auditing, testing and evaluation, explainability, and specifically how to adjudicate Equal Protection Clause or Title VII cases. Analysis shows that these recommendations are often just as narrow, overconfident, or incompatible with our cybernetic reality as the underlying legal system itself.

Subsection V.B. acknowledges that our process-based anticlassification beliefs are preventing us from addressing the proper antisubordination goal of people to be free from discriminatory outcomes. Therefore, I argue that discriminatory outcomes must be enforced through strict liability where intent is presumed once discriminatory outcomes are identified. Strict liability is necessary when those harmed cannot adequately enforce the obligations necessary to ensure reasonable quality control,[358] and everything in this Essay has shown that that is the case with antidiscrimination: plaintiffs cannot adequately enforce protections against cybernetic system discrimination. Moreover, because this puts all the pressure on the question of how to identify discriminatory outcomes, it demands the necessary antisubordination conversation that is long overdue in our society. We have to ask ourselves: what amount and types of "discriminatory

[356] FAIRNESS IN EMPLOYMENT TESTING, *supra* note 321.

[357] Amy L Wax, *Disparate Impact Realism*, 53 WILLIAM MARY LAW REV. 621, 654 (2011).

[358] Geistfeld, *supra* note 59 at 1664. Strict liability also compelled from a moral perspective to ensure anti-discrimination law is achieves accountability. Wirts, *supra* note 19 at 849–50.

misperformance" are we ok with?

A. *Current Recommendations are Important but Insufficient*

There are three main groups of recommendations addressing cybernetic system discrimination: The first group argues for fairness, access, auditing, testing and evaluation, and financial and technical expertise addressing discrimination arising from inputs and the machine. The second advocates for explainability and testing and evaluation to address discrimination arising from human-machine interaction, users, and feedback. The third focuses on the law, arguing for updating the rules regarding the Equal Protection Clause or Title VII by increasing types of evidence accepted by courts or encouraging A/B testing can better identify discrimination, particularly machine discrimination. By integrating all these recommendations two results become immediately clear: first, there are serious technical difficulties in implementing any of these recommendations, and two, although many of these recommendations are valuable, no one solution can alone seriously address the depths and varieties of conflicts between the assumptions in current antidiscrimination law and the realities of cybernetic system discrimination as identified in this Essay.

a. Inputs and the Machine

Perhaps the most popular recommendations in legal scholarship today are those attempting to find technological solutions to machine discrimination, namely fixing the inputs and the machine. This section will show that the proposals can help identify and reduce discrimination but only if they are jointly addressed by scientists and engineers, courts, and governments: collectively agreed-upon definitions for machine fairness (or how machines should be designed), meaningful access for stakeholders and independent third-parties to the machines, new state-of-the-art methods for test and evaluation, and the financial and technical ability for those most often affected by discrimination to enforce their civil and human rights. However, these proposals, despite their popularity, suffer a range of issues: focusing on machine discrimination while ignoring the reality that the machine is only one component of the full cybernetic system; adopting a techno-solutionist perspective by trying to solve normative (social and moral) issues with technology; overstating the likelihood that these technical solutions can or will be adopted; and entrenching a world where colorblindness is the norm instead of substantive justice.

To begin, Kroll and his computer science colleagues called for the design of "accountable algorithms" building in software verification, zero-knowledge proofs, cryptographic commitments, and fair random choices to ensure "procedural regularity."[359] The goal here is for an algorithm designer to ensure that a "particular algorithm does not directly use sensitive or prohibited classes of information, such as gender, race, religion, or medical status."[360] From a technical perspective, they acknowledge that the ability of machine learning algorithms to adapt and change may need even further technical solutions like incorporating randomness to maximize the algorithm's learning, a method for maximizing fairness, and differential privacy.[361] Like

[359] Joshua Kroll et al., *Accountable Algorithms*, 165 UNIV. PA. LAW REV. 633–705, 662.

[360] Kroll et al., *supra* note 41 at 682.

[361] *Id.* at 683.

turtles, it seems to be technology all the way down – discrimination can be solved with flawed technology, whose flaws can be solved with different flawed technology, and on and on.

Belying these layers of technical solutions, Kroll and his colleagues correctly state that any of these technical solutions must be in service of the normative goals established outside of the designer's control.[362] This reality is best exemplified by the concept of machine or algorithmic fairness.[363] Talia Gillis summarizes the many approaches to "algorithmic fairness" into four groups: excluding the protected class characteristics, excluding the proxies for protected class characteristics, restricting the inputs to pre-approved characteristics, and orthogonalizing inputs in order to prevent bias from omitted variables.[364] But as she empirically showed, this "embrace [of] the input-centered approach of traditional law… fail on their own terms, are likely unfeasible, and overlook the benefits of accurate prediction."[365] So focusing on the inputs seems to be insufficient.

Deeper than the technical and empirical concerns, the fundamental critique of algorithmic fairness is that we should not and cannot "equate technical and social notions of fairness… reduce[ing fairness] to a single mathematical definition that exists in the abstract, apart from social, political, and historical context."[366] The value judgements of fairness cannot come from technologists or data scientists but must come from domain experts and affected populations.[367] Basing conclusions of fairness on training data and lists of legally protected categories without addressing social contexts and the philosophies of fairness and justice "misunderstand[s]… [and] inappropriately co-opt[s] legal language,"[368] creating immense practical challenges that limit the effectiveness and consistency of any fairness mechanism.[369] The "myopia of 'fairness'"[370] tends to ignore intersectionality and the broader contexts of discrimination,[371] especially when

[362] *Id.* at 678. ("Technical tools offer ways to ameliorate these problems, but they generally require a well-defined notion of what sort of fairness they are supposed to be enforcing.")

[363] Kroll et al., *supra* note 360. ("We emphasize that computer scientists cannot assume that the policy process will give them a meaningful, universal, and self-consistent theory of fairness to use as a specification for algorithms. There are structural, political, and jurisprudential reasons why no such theory exists today. Likewise, the policy process would likely not accept such a theory if it were generated by computer scientists.")

[364] Talia B. Gillis, *False Dreams of Algorithmic Fairness: The Case of Credit Pricing*, SSRN ELECTRON. J., 44 (2020), https://www.ssrn.com/abstract=3571266 (last visited Mar 21, 2021). Rashida Richardson, Jason Schultz & Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N. Y. UNIV. LAW REV. 192–233, 224 (2019). ("Thus, restrictions or prohibitions on the use of the historical data generated by unlawful and biased practices are necessary to ensure that the legacy of such practices is not perpetuated through the systems that rely on such data.")

[365] Gillis, *supra* note 365 at 2.

[366] Ben Green, *The false promise of risk assessments: epistemic reform and the limits of fairness, in* PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 594–606, 10 (2020), https://dl.acm.org/doi/10.1145/3351095.3372869 (last visited Apr 12, 2021).

[367] Sorelle A. Friedler, Carlos Scheidegger & Suresh Venkatasubramanian, *The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making*, 64 COMMUN. ACM 136–143, 143 (2021).

[368] Alice Xiang & Inioluwa Deborah Raji, *On the Legal Compatibility of Fairness Definitions*, ARXIV191200761 CS STAT (2019), http://arxiv.org/abs/1912.00761 (last visited Sep 25, 2020).

[369] Reuben Binns, *Fairness in Machine Learning: Lessons from Political Philosophy*, 81 PROC. MACH. LEARN. RES. 1, 9 (2018).

[370] Green, *supra* note 367 at 10.

[371] Hoffmann, *supra* note 116.

implemented by machines or algorithms intentionally devoid of underlying theories grounded in people's experiences of discrimination.[372]

As called for in the literature, and exemplified throughout this Essay, the answer to the myopia of fairness or any number of techno-solutionist proposals is to embrace the reality that these technologies exist within complex, interdependent cybernetic systems where the inputs and machine are but a small part. As Green explains, we must "approach[] algorithms as sociotechnical imaginaries rather than as discrete technologies… . By highlighting the entire context surrounding algorithms as subject to reimagination and reform, this approach avoids the trap of false dilemmas and makes possible more substantive change." [373] Even notionally "fairness-aware" algorithms can be ineffective, inaccurate, or dangerously misguided when they do not account for the realities of cybernetic systems.[374] I think it is fair to assume that plaintiffs would prefer a judge who appreciated the complexities of plaintiff's lives and real effects of discrimination rather than just someone with a Ph.D. in artificial intelligence.

But assuming we can somehow agree on normatively fair methods and measures for the fair design of algorithms, there is the next critical question: how can third parties like plaintiffs alleging discrimination or courts themselves be assured that the machines were designed and operating effectively post-deployment to achieve these normative measures of fairness? Here again, researchers have identified two further serious complications to holding cybernetic systems accountable: lack of access and the difficulty of meaningfully testing and evaluating these systems.

Many machines are subject to contracts from private vendors and broadly protected by trade secrets and intellectual property rights preventing meaningful access.[375] Rebecca Wexler comprehensively detailed the rise of a trade secret privilege in criminal proceedings preventing defendants from meaningful access to the technologies informing and often determining the outcome of their case.[376] But intellectual property has become a barrier to transparency and accountability in almost any aspect of our public and private domains, from public infrastructure and commercial activities to healthcare and administrative rulemaking.[377] Separately, various jurisdictions have interpreted the Computer Fraud and Abuse Act and the Digital Millennium Copyright Act to prevent third parties from evaluating machines for potential discrimination.[378]

[372] Konstantinos V. Katsikopoulos & Marc C. Canellas, *Decoding Human Behavior with Big Data? Critical, Constructive Input from the Decision Sciences*, AI MAG., 13 (2021).

[373] Green, *supra* note 367 at 10.

[374] Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, *in* PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY - FAT* '19 59 (2019), http://dl.acm.org/citation.cfm?doid=3287560.3287598 (last visited Sep 25, 2020).

[375] Hannah Bloch-Wehba, *Access to Algorithms*, 88 FORDHAM LAW REV. 1265 (2020).

[376] Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STANFORD LAW REV. 1343–1429 (2018).

[377] *Id.* at 1351 nn. 30–34. (collecting references). See also, Sonia K Katyal, *The Paradox of Source Code Secrecy*, 104 CORNELL LAW REV. 98 (2020).

[378] PETER STONE ET AL., *Artificial Intelligence and Life in 2030* 43 (2016), https://ai100.stanford.edu/2016-report. (Recommending that we "[r]emove the perceived and actual impediments to research on the fairness, security, privacy, and social impacts of AI systems. Some interpretations of federal laws such as the Computer Fraud and Abuse Act and the anti-circumvention provision of the Digital Millennium Copyright Act are ambiguous regarding whether and how proprietary AI systems may be reverse engineered and evaluated by academics, journalists, and other researchers. Such research is critical if AI systems with physical and other material consequences are to be properly vetted and held accountable.")

Authors describe these intellectual property barriers as paradoxes[379] or excessive[380] but they are still holding strong. First, there is still a long path toward reforming the laws around transparency and disclosure,[381] open data,[382] or using freedom of information and access laws.[383] Second, these solutions are also technologically myopic. Yes, the machines, the algorithms, and the data are important, but no amount of transparency into those elements will provide transparency into the other aspects determining the outputs of cybernetic systems. How does transparency into the machine help identify a potential failure in human-machine interaction or access to the source code help identify a failed feedback loop? This is why Ananny and Crawford ultimately conclude that the current construct of transparency "is an inadequate way to understand—much less govern—algorithms."[384]

But, once again, let us assume that we have solved the problems above of fairness and now access. We have complete control of the machine, but now we are left with the question of how to meaningfully test and evaluate the inputs and machine to determine if these systems are failing, causing discriminatory outcomes. That is, ultimately the reason for the access: transparency in order to audit and challenge evidence and decisions. The most popular term for this in the legal literature is auditing.[385] As defined by Kroll *et al.*, "[i]n computer science, auditing refers to an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures. Auditing is intended to reveal whether the appropriate procedures were followed and to uncover any tampering with a computer system's operation."[386] This could be audit trails that record facts and bases of the system's decisions,[387] examining inputs and outputs to discover problems,[388] or providing the machines simulated or mock data to analyze outcomes.[389]

Despite the popularity of calling for auditing algorithms, there are serious process and philosophical concerns difficulties. The narrowness of auditing alone is exemplified by the description of the popular algorithmic impact assessments (or statements) which include auditing but only as one element alongside broader cybernetic needs like self-assessments, external researcher reviews, public notice, and due process mechanisms.[390] For employment discrimination under Title VII, *Ricci* relegates auditing to pre-deployment. Pragmatically, there is also the growing practice of developers and organizations purchasing and hijacking audits to limit the

[379] Katyal, *supra* note 378.

[380] Jeanne C Fromer, *Machines as the new Oompa-Loompas: trade secrecy, the cloud, machine learning, and automation*, 94 N. Y. UNIV. LAW REV. 706 (2019).

[381] Katyal, *supra* note 378.

[382] Cynthia Conti-Cook, *Open Data Policing*, 106 GEORGET. LAW J. ONLINE 1–23 (2017).

[383] Bloch-Wehba, *supra* note 376.

[384] Ananny and Crawford, *supra* note 103 at 974.

[385] Kim, *supra* note 40.

[386] Kroll et al., *supra* note 41 at 660–61. (quoting IEEE Computer Society, IEEE Std 1028 – IEEE Standard for Software Reviews and Audits §8.1 (Aug. 15, 2008)).

[387] See e.g., Citron and Pasquale, *supra* note 40 at 28; Danielle Citron, *Technological Due Process*, 85 WASH. UNIV. LAW REV. 1249–1313, 1277 (2008).

[388] Kim, *supra* note 40 at 190.

[389] Inioluwa Deborah Raji & Joy Buolamwini, *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*, *in* PROCEEDINGS OF THE 2019 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 429 (2019), https://dl.acm.org/doi/10.1145/3306618.3314244 (last visited Sep 25, 2020).

[390] DILLON REISMAN ET AL., *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability* 4 (2018). See also, Barocas and Selbst, *supra* note 13; Selbst, *supra* note 42.

capabilities of the auditor and labeling the result as "collaborative" (non-independent) auditing.[391] As Mona Sloane warns, if we "grac[e] such technologies with an audit that is steered by the organization being audited and only examine[] if an algorithm 'works as intended,' researchers and technologists alike become complicit in legitimizing and normalizing weak and problematic notions of algorithmic auditing, as well as technologies that, by their very definition, are discriminatory."[392]

Even with all of these challenges, the most difficult and important part of this whole exercise is the actual testing and evaluating the machine to determine if it is biased or discriminatory. As shown in this Essay with the firefighter officer exam or the GATB, even machines without the buzzword components of artificial intelligence require immense effort to determine the source of their discrimination. Turning to the modern machines increasingly infused with artificial intelligence and autonomy one finds even more challenges. These machines can be as "opaque as the brain" with experts lamenting that despite 25 years of development, "deciphering the black box has become exponentially harder and more urgent. The technology itself has exploded in complexity… ."[393] Remember, as described above in Section III, duality, deconstruction, and structuralism are myths. As a result, "[a]pproaches that attempt to review system failures simply by looking at how the output responds to changes in input are limited by either an inability to attribute a cause to those changes or an inability to interpret whether or why a change is significant."[394] This is why Kroll *et al.* note that use access to the source code and audits are "not sufficient to provide accountability in all cases" [395] because many systems are not designed with evaluation in mind. These machines will incorporate randomness into their processes making outcomes potentially unpredictable and unreproducible. This means that some information determining machine performance will never be discoverable as they never take "durable, observable forms[. F]or example, an 'algorithm could compute a variable in memory that corresponds to some protected class such as race,' but if the memory exists only temporarily 'FOIA would be unable to compel its disclosure.'"[396] There are also issues where the machine has been attacked or manipulated by third parties in imperceptible ways[397] or is being repeatedly updated, requiring repeated testing and evaluation for new issues.[398]

[391] Mona Sloane, *The Algorithmic Auditing Trap*, MEDIUM ONEZERO (2021), https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d (last visited Apr 21, 2021).

[392] *Id.*

[393] Davide Castelvecchi, *Can we open the black box of AI?*, 538 NATURE 21, 21 (2016).

[394] Kroll et al., *supra* note 41 at 661. ("For example, there is a substantial body of literature in computer science that addresses audits of electronic voting systems, and security experts generally agree that proper auditing is necessary but insufficient to assure secure computer-aided voting systems. Computer scientists, however, have shown that black-box evaluation of systems is the least powerful of a set of available methods for understanding and verifying system behavior. Even for measuring demonstrable properties of software systems, such as testing whether a system functions as desired without bugs, it is much more powerful to be able to understand the design of that system and test it in smaller, simpler pieces. Approaches that attempt to review system failures simply by looking at how the output responds to changes in input are limited by either an inability to attribute a cause to those changes or an inability to interpret whether or why a change is significant.") (citations omitted)

[395] Kroll et al., *supra* note 360 at 657.

[396] Ananny and Crawford, *supra* note 103 at 981. (quoting K. Diakopoulos, *Accountability in algorithmic decision making*, 59 COMMUN. ACM 56, 59 (2016).)

[397] *See e.g.*, Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* (2018).

[398] See discussions over STRmix version numbers.

To fully understand the immense challenges ahead, there is no better place to look than the professionals responsible for testing and evaluating many of the most advanced military and defense systems in the world. Professional testers explain the numerous practical barriers including the lack of built-in (not ex-post) transparency, the lack of mechanisms for recording data, and the difficulty of testing for emergent behavior.[399] As Felder and Collopy explain for cybernetic systems: "The combination of many component systems (with autonomy, diversity, dynamic connectivity and belonging) and component systems with non-trivial internal complexity (which generates emergence) results in an almost infinite number of possible states for a typical system of systems."[400] For example, to fully test new software responsible only for automatically delivering flight information to an aircraft parked at a gate, "will be literally impossible" using current techniques.[401]

In closing, even if we just focus on the machine failure and discrimination alone, we will need to develop an agreed-upon mechanism for machine fairness (or how they should be designed), stakeholders must be given access for auditing, and the practical barriers of transparency, design, statistics and modeling must be solved. This is a complex mix of barriers that require significant advancements by technologists, courts, and governments alike.

But let us imagine that we have solved all the above. Even still last challenge that should not be forgotten: almost none of the people most affected by these systems will be able to utilize these new avenues to ensure the machines are adhering to the normative rules. Governments can barely keep up with governance and oversight over technology,[402] let alone the attorneys on the front lines responsible for litigating and enforcing constitutional, civil and human rights – assuming the plaintiff even has access to an attorney. For example, by almost any analysis, public defenders representing those accused of crimes but unable to afford an attorney are themselves incredibly underfunded and overworked.[403] Many defenders are barely able to ensure basic Constitutional rights, let alone have the resources, time, and expertise to litigate potentially discriminatory

[399] DANIEL PORTER ET AL., *Trustworthy Autonomy: A Roadmap to Assurance, Part 1: System Effectiveness* (2020).

[400] Felder and Collopy, *supra* note 189 at 324.

[401] *Id.* at 324.

[402] 7 GARY E MARCHANT, BRADEN R ALLENBY & JOSEPH R HERKERT, THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND LEGAL-ETHICAL OVERSIGHT: THE PACING PROBLEM (2011).

[403] *See, e.g.,* AM. BAR ASS'N STANDING COMMITTEE ON LEGAL AID AND INDIGENT DEFENDANTS, GIDEON'S BROKEN PROMISE: AMERICA'S CONTINUING QUEST FOR EQUAL JUSTICE iv (December 2004) ("Overall, our hearings support the disturbing conclusion that thousands of persons are processed through America's courts every year either with no lawyer at all or with a lawyer who does not have the time, resources, or in some cases the inclination to provide effective representation. All too often, defendants plead guilty, even if they are innocent, without really understanding their legal rights or what is occurring. Sometimes the proceedings reflect little or no recognition that the accused is mentally ill or does not adequately understand English. The fundamental right to a lawyer that Americans assume apply to everyone accused of criminal conduct effectively does not exist in practice for countless people across the United States."); NAT'L ASS'N OF CRIM. DEF. LAW.'S, MINOR CRIMES, MASSIVE WASTE: THE TERRIBLE TOLL OF AMERICA'S BROKEN MISDEMEANOR COURTS 7 (April 2009) ("Legal representation for misdemeanants is absent in many cases. When an attorney is provided, crushing workloads often make it impossible for the defender to effectively represent her clients. Counsel is unable to spend adequate time on each of her cases, and often lacks necessary resources, such as access to investigators, experts, and online research tools. These deficiencies force even the most competent and dedicated attorneys to engage in breaches of professional duties. Too often, judges and prosecutors are complicit in these breaches, pushing defenders and defendants to take action with limited time and knowledge of their cases. This leads to guilty pleas by the innocent, inappropriate sentences, and wrongful incarceration, all at taxpayer expense.").

machines.[404] Moreover, the ability to access, assess, and litigate these machines requires experts[405] which few offices can afford and who are often forbidden from working with defenders.[406]

b. Human-Machine Interaction

Beyond all of these substantial technical, policy, and legal barriers to addressing cybernetic system discrimination in the inputs and machine, there is still the human-machine interaction, user, and feedback to address – in other words, the rest of the cybernetic system. It is essentially to remember that no matter the amount of artificial intelligence, statistics, or mathematics, there will always be a human user somewhere in the deployment of a cybernetic system:[407] hiring software are used by hiring managers; risk assessment tools are used by judges or child welfare representatives; and, DNA software and image recognition systems are used by forensic analysts and prosecutors. We should not "permit[] ignorance of the ways humans and technology co-conspire to not just passively reproduce but actively uphold and reproduce discriminatory social structures… ."[408] As Ananny and Crawford explain, "[a]n algorithmic system is not just code and data but an assemblage of human and non-human actors—of institutionally situated code, practices, and norms with the power to create, sustain, and signify relationships among people and data through minimally observable, semiautonomous action. This requires going beyond algorithms as fetishized objects to take better account of the human scenes where algorithms, code, and platforms intersect."[409] As Selbst and Barocas point out, to explain the machine's influence on the ultimate decision requires more than just a complete, documented explanation of the model, it requires a complete, documented explanation of how the model interacts with the human and the organization within the cybernetic system: "Models are not self-executing; an additional layer of decisions concerns the institutional process that surrounds the model. Are the model outputs automatically accepted as the ultimate decisions? If not, how central is the model to the decision? How do decision makers integrate the model into their larger decision frameworks? How are they trained to do so? What role does discretion play?" [410] To respond to these questions, this section opens by describing the promise and difficulties of building "explainable" systems, and concludes with the difficulties of testing and evaluating human-machine interaction and feedback in cybernetic systems.

[404] *Id.*

[405] Kirchner, *supra* note 38.

[406] INVESTIGATION AND EXPERTS: A GUIDE TO THE IMPLEMENTATION OF THE MINIMUM STANDARDS FOR DELIVERY SYSTEMS, (2017), https://michiganidc.gov/wp-content/uploads/2017/03/White-Paper-3-Experts-and-Investigators.pdf. Kashmir Hill, *Imagine Being on Trial. With Exonerating Evidence Trapped on Your Phone.*, NEW YORK TIMES, November 24, 2019, at 1. ("It's definitely an uneven playing field. Law enforcement has an understandable desire to extricate data from the digital world to solve cases, but there hasn't been adequate scrutiny of these new techniques. Law enforcement agencies get a new investigative technique — fingerprinting, DNA analysis, breathalyzer tests — and those representing the accused struggle to play catch-up. Developing the new technical expertise necessary to adequately defend their clients is a challenge. Not only do public defenders tend to be underfunded, law enforcement can monopolize the experts in the field and forbid them from working for the defense.") (internal quotation marks omitted).

[407] Jon Kleinberg et al., *Discrimination in the age of algorithms*, 2010 J. LEG. ANAL. 113, 163 (2019). Meg Leta (Ambrose) Jones, *The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practices Principles*, 18 VANDERBILT J. ENTERTAIN. TECHNOL. LAW 77 (2015).

[408] Hoffmann, *supra* note 116 at 904.

[409] Ananny and Crawford, *supra* note 103 at 983. (internal citations and quotes omitted).

[410] Andrew Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM LAW REV. 1085, 1132 (2018).

Within the legal research on anti-discrimination relating to human-machine interaction, there has been an almost exclusive focused on the transparency of the machine to the human user, or "explainability."[411] The vast drive towards explainable systems is motivated by how difficult modern machines are to understand, even to those who design and build the system.[412] Therefore, explainability is suggested as a key means to engender trust, to help understand the causal relationships in the machine, or to achieve the legal right to explanation.[413] Selbst and Barocas summarized the three motivations as an explanation being an inherent good, necessary for respecting autonomy, dignity, and personhood; explanation enabling action such as those necessary for legal accountability; and as a mechanism for evaluating the validity and justifiability of making decisions based on the machine.[414] Legal requirements for explainability exist not only in the famous European Union's General Data Protection Regulation (GDPR) Article 13-15 but also implicitly in the Fair Credit Reporting Act (FCRA) and the Equal Credit Opportunity Act (ECOA).[415]

There are both philosophical and technical barriers to achieving systems that are explainable to their users. First, some researchers have questioned how useful the concept of explainability would be as a remedy or prevention for algorithmic harms.[416] Maybe rather than a "right to an explanation" such as that in the GDPR, we should demand a right to good decisions.[417] (Hinting again at the difference between anticlassification's focus on process and antisubordination's focus on outcomes.) As comprehensively summarized by Ananny and Crawford, the broader ideal of explainability and transparency could have shortcomings in implementation including not necessarily building trust,[418] privileging seeing over understanding, and a failure to attend to the technical and temporal limitations.[419]

Even if society agrees on the value of explainability for users, it must address the technical limitations which have caused some researchers to describe explainability as part of a "transparency fallacy."[420] Explainability has become a buzzword without a solid foundation that

[411] Here I will use the term explainability but many of the articles use the word interpretability interchangeably.

[412] Julie Gerlings, Arisa Shollo & Ioanna Constantiou, *Reviewing the Need for Explainable Artificial Intelligence (xAI)*1284 (2021), http://hdl.handle.net/10125/70768 (last visited Aug 9, 2021); Randy Goebel et al., *Explainable AI: The New 42?*, 11015 *in* MACHINE LEARNING AND KNOWLEDGE EXTRACTION 295 (2018), http://link.springer.com/10.1007/978-3-319-99740-7_21 (last visited Aug 9, 2021).

[413] Zachary Lipton, *The Mythos of Model Interpretability*, 16 ACM QUEUE 1–27, 4–5 (2018). (internal citations omitted). See also, Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 UNIV. CHIC. LAW REV. 1311, 1366 (2019). (challenging our intuitions as to why we may want explanations).

[414] Selbst and Barocas, *supra* note 411 at 1118–26.

[415] *Id.* at 1099–100. See also, Tim Miller, *Explanation in Artificial Intelligence: Insights from the Social Sciences*, ARXIV170607269 CS (2018), http://arxiv.org/abs/1706.07269 (last visited Apr 23, 2021).

[416] Cynthia Dwork & Deirdre K Mulligan, *It's Not Privacy, and It's Not Fair*, 66 STANF. LAW REV. ONLINE 35 (2013).

[417] Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For*, 16 DUKE LAW TECHNOL. REV. 18–84 (2017); Lilian Edwards & Michael Veale, *Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?*, IEEE SECUR. PRIV. 46–54 (2018).

[418] Canellas et al., *supra* note 44 at 22–27.

[419] Ananny and Crawford, *supra* note 103.

[420] Edwards and Veale, *supra* note 418 at 65.

now consists of numerous, often conflicting, motivations, definitions and techniques.[421] The true and accurate explanation for a machine's decision may be simply beyond our intuition. The machine's decisions or processes "might not even lend themselves to hypotheses about what accounts for the models' discoveries. Parsimonious models lend themselves to more intuitive reasoning, but they have limits—a complex world may require complex models."[422] In practice, many "explanations" are provided not from the machine making the decision, but from a second, new machine that attempts to approximate the predictions of the first machine in a way that is easier to explain to the user.[423] Imagine a news reporter explaining why a jury came to a verdict without having been in the room or speaking with any of the jurors. As a result, some purported explanations are completely decoupled from machine actually making the decision or recommendation. Therefore, professional data scientists "often have trouble understanding the exact relationship between the two models or explaining this relationship to stakeholders [(users)], which is ultimately undermining the value of both models and the whole enterprise [of transparency and explainability]."[424]

Stepping outside of explainability and returning to the perspectives of professional testers, the already substantial difficulties of testing and evaluating modern machines discussed above are only amplified when considering the full cybernetic system of machines and humans interacting within organizations. In truth, while testing cybernetic systems is not new, it is still a small domain attempting to grow and adapt to the exponentially growing number and complexity of these systems. It is much easier to build and deploy a cybernetic system than it is to show with confidence that it is safe, reliable, or non-discriminatory. As a result, unsolved challenges litter the field. For example, most failures will be incredibly rare and the result of emergent behavior making it difficult to design test environments to generate emergent behavior,[425] especially given that humans are adept at resolving abnormalities before deeper failures can be revealed.[426] Moreover, many machines are explicitly designed to change their capabilities, functionality, and interaction with humans over time requiring the human user to also adapt to the machine over time in order to maintain performance. Therefore, auditing or certifying a machine or cybernetic system pre-deployment is insufficient to meaningfully characterize future performance despite what the *Ricci* majority and other commentators may say.[427] Even fundamental models and measures like basic

[421] Lipton, *supra* note 414.

[422] Selbst and Barocas, *supra* note 411 at 1129.

[423] Scott M Lundberg & Su-In Lee, *A Unified Approach to Interpreting Model Predictions*, *in* PROCEEDINGS OF THE 31ST CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS 2017) 1 (2017).

[424] Katsikopoulos and Canellas, *supra* note 373 at 14. (citing Harmanpreet Kaur et al., *Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning*, *in* PROCEEDINGS OF THE 2020 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1 (2020), https://dl.acm.org/doi/10.1145/3313831.3376219 (last visited Apr 23, 2021); Samir Passi & Steven J Jackson, *Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects*, 2 PROC. ACM HUM.-COMPUT. INTERACT. 1 (2018); I. Elizabeth Kumar et al., *Problems with Shapley-value-based explanations as feature importance measures*, ARXIV200211097 CS STAT (2020), http://arxiv.org/abs/2002.11097 (last visited Apr 23, 2021).)

[425] PORTER ET AL., *supra* note 400 at iii. See also, Canellas et al., *supra* note 44 at 40. ("Computational[] simulations have to account for an extraordinary range of interactions in which the events are rare and systematic recorded data even rarer, making duplicating exact circumstances difficult") (internal citations omitted)

[426] Pritchett, *supra* note 190.

[427] Kim, *supra* note 40.

statistics are having to be rethought and redefined to quantify and predict performance of cybernetic systems which often have high rates of normal performance coupled with high rates of catastrophic performance.[428] In closing, we must remember that if professional testers and researchers are still attempting to understand the scope of testing and evaluating cybernetic systems, there is no reason to expect a lay person, a public defender, or someone without extensive means and time to meaningful do so.

c. Equal Protection Clause and Title VII

Turning from the element-by-element analysis of proposed solutions, this subsection briefly examines some of the specific proposals for adjusting the laws around the Equal Protection Clause and Title VII. First, Huq has recommended that courts could restore the types of evidence first highlighted in *Arlington Heights*[429] – semantic context, the statements of officials, political context, depositions and interrogatories, and statistical evidence – but it is not clear how any of those five types of evidence will resolve the numerous barriers described above. Scherer *et al.* proposed a unified disparate impact framework but again centers entirely on the algorithms, keeping the business necessity defense reliant on the outdated conceptions of validity in the EEOC Guidelines and retaining the plaintiff's impossible burden of proving a less discriminatory alternative.[430] Bathaee suggests reconstructing the intent test based on how "autonomous" and "transparent" the system is[431] but there is no meaningful ordinal scale or mechanism for easy comparison of systems based on autonomy[432] or transparency.[433]

Kim emphasized the belief that *Ricci* is not an impediment to reducing machine discrimination in employment decisions as there is "nothing in Ricci prevents a court from enjoining the use of a biased model, or an employer from voluntarily ceasing to use the discriminatory algorithm once that bias has been detected."[434] From the discussions above, we already have seen numerous problems with this view of addressing cybernetic discrimination. This is algorithm-centric, ignoring the issues of how the algorithm is used or interpreted by the human user within the context of their sociotechnical system. It is incredibly difficult to determine ex-ante whether an algorithm will be discriminatory in operation. Only requiring an employer to stop using an algorithm after harm has been caused and without any civil penalties, does not do much to encourage a true

[428] Amy R Pritchett, So Young Kim & Karen M Feigh, *Modeling Human–Automation Function Allocation*, 8 J. COGN. ENG. DECIS. MAK. 33–51 (2014); Pritchett, Kim, and Feigh, *supra* note 212; Canellas et al., *supra* note 44 at 40–41. Felder and Collopy, *supra* note 189.

[429] Huq, *supra* note 236.

[430] Scherer, King, and Mrkonich, *supra* note 307 at 500.

[431] Bathaee, *supra* note 42.

[432] "[D]espite their former prevalence in the academic literature, [single-dimensional levels of autonomy] is now acknowledged to be limited, problematic, and, to some, worth discarding altogether." Canellas and Haga, *supra* note 47 at 32. (citing Bradshaw et al., *supra* note 129; Karen M Feigh & Amy R Pritchett, *Requirements for Effective Function Allocation A Critical Review*, 8 J. COGN. ENG. DECIS. MAK. 23–32 (2014).)

[433] Canellas et al., *supra* note 44 at 25. (summarizing the various models of transparency with purposes including trust, fault finding, or validation and verification).

[434] Kim, *supra* note 40 at 931–32. (critiquing Barocas and Selbst for "erroneously suggest[ing] that *Ricci* poses an obstacle to crafting a remedy for biased classification schemes") (citing Barocas and Selbst, *supra* note 56 at 725–26.)

rethinking of how to prevent harms. Then, as Bent points out, "[t]here is a significant difference between *discarding* a biased algorithm and *fixing* a biased algorithm by introducing a race-aware fairness constraint."[435] While Bent is focused here on the need to use race information to adjust the algorithm, there is a separate important aspect: how can courts or employees be confident that a previously-biased algorithm is "fixed" when the employer was not able to ensure it was biased before? More cynically, what is to stop a developer or employer from taking an algorithm labeled as biased, change a few lines of code, promise it is "fixed," sell it or use it for as long as possible, be called out for bias, then rinse and repeat.

Bent also proposes that the "but-for causation" as required for showing discrimination under the ADEA "might prove challenging for real-world litigants, but in theory it should not be difficult at all."[436] Putting aside the glossing over of the needs of real-world litigants, who one may naïvely think should be the focus of real-world legal scholars, how would this work in theory? Bent proposes A/B testing where:

> "Programmers can delete the fairness constraint instructions and leave the program with only one optimization instruction: 'pick good employees.' Then the results for any individual candidate could be directly compared, with and without the fairness constraint. A plaintiff might have been classified as 'bad employee—don't hire' using an algorithm with the race-aware fairness constraint, but classified as "good employee— hire" using the same algorithm without the fairness constraint. If so, the plaintiff would have an unusually strong case that the race-aware fairness constraint was a but-for cause of the adverse employment action." [437]

This proposal is flawed for numerous reasons but just to mention three: the idea that there is an optimization instruction merely saying "pick good employees," that aspects of software can simply be "deleted" (or "commented out") without any cascading effects on the operation of the system, and the narrow focus on the algorithmic fairness despite the many other complications of cybernetic systems.

B. *Our Only Option is to Enforce Discriminatory Intent via Strict Liability and Anti-Subordination*

The above set of legal proposals, much like the element-by-element recommendations before it, seems unable to comprehend and therefore address the full breadth of ways that cybernetic systems can fail and discrimination. Even if we achieve the near-impossible goals of removing human discrimination, machine discrimination, and systemic discrimination, we will still have cybernetic system discrimination. This leaves us still searching. Maybe the problem is that we have focused too much on the process instead of the outcomes. Like the colorblindness principle, we may have become too focused on process, instead of the antisubordination principle's

[435] Bent, *supra* note 42 at 833.

[436] *Id.* at 829.

[437] *Id.* at 829. (internal citation omitted) *See also, Id.* at 829 fn. 127. ("This could be done by 'commenting out' the part of the computer code that implements the fairness constraint, which would convert that portion of code from an instruction to a comment that the computer ignores.")

desire to focus on outcomes. The reason for vast research dedicated to fairness, access, auditing, explainability, and testing and evaluating, is because of concern over the discrimination present in our society. So rather than prioritize a right to fair algorithms, access, audits, explainability, testing and evaluation, new forms of evidence, we should instead prioritize a right to be free from discriminatory outcomes and view fair algorithms, access, audits, explainability, testing and evaluation, new forms of evidence, or any other proposal as a means to that end. This section provides a path toward this goal strict liability and antisubordination.

Following in the footsteps of Charles Lawrence, I do not set out to invent a better tool for anti-discrimination litigators.[438] I seek to challenge the paradigm of anti-classification itself. Yes, this Essay has shown that when we use machines in our decision making we become a part of cybernetic systems with numerous sources of potential failure and discrimination,[439] that these cybernetic systems are inherently complex and interdependent,[440] which our antidiscrimination laws are incapable of understanding[441] and which no amount of popular proposals can fully address.[442] But this is all the foundation for this Essay's true purpose: a rejection of the anticlassification paradigm as the sole purpose of antidiscrimination law.

The anticlassification principle holds that "the responsibility of law is to eliminate the unfairness in certain protected classes experience due to decision makers' choices."[443] Therefore, "where the internal difficulties cannot be overcome, there is likely no way to correct for the discriminatory outcomes…" because "as long as employers are not intentionally discriminating based on explicitly proscribed criteria, the chips should fall where they may."[444] But from any legitimate understanding of how cybernetic systems operate in reality, focusing solely on decision makers' choices will not eliminate unfairness and there is almost no way that these internal difficulties can be truly overcome. So, to continue to believe in anticlassification as the sole purpose of antidiscrimination law in the face of cybernetic systems is to abandon the belief in the law truly remedying discrimination.

But there is another way. The antisubordination principle "holds that the goal of antidiscrimination law is, or at least should be, to eliminate status-based inequality due to membership in those classes, not as a matter of procedure, but of substance."[445] Antisubordination

[438] Charles Lawrence III, *Unconscious Racism Revisited: Reflections on the Impact and Origins of "The Id, the Ego, and Equal Protection"*, 40 CONN. LAW REV. 931–978, 964 (2010). (Reflecting on Lawrence, *supra* note 19."I did not set out to invent a better tool for Title VII litigators. I sought to challenge the disparate treatment paradigm itself. I argued that racism's harm was greater than the biased actions of individuals. I pointed to the ubiquity of conscious and unconscious racism as evidence of the continued vitality of racist ideology and argued that so long as this ideology lived and flourished, the Constitution, and normative justice, required that we act affirmatively to remedy its effects and disestablish its institutional embodiments.")

[439] *Supra* Section **Error! Reference source not found.**

[440] *Supra* Section III.B.

[441] *Supra* Section IV.

[442] *Supra* Section V.A.

[443] Barocas and Selbst, *supra* note 56 at 723. (internal citations omitted) Norton, *supra* note 13. See also, Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition-- Anticlassification or Antisubordination?,* 58 UNIV. MIAMI LAW REV. 9 (2003); Reva B. Siegel, *Equality Talk: Antisubordination and Anticlassification Values in Constitutional Struggles over Brown*, 117 HARV. LAW REV. 1470 (2004); Ruth Colker, *Anti-Subordination Above All: Sex, Race, and Equal Protection*, 61 N. Y. UNIV. LAW REV. 1003 (1986).

[444] Barocas and Selbst, *supra* note 13 at 726.

[445] Barocas and Selbst, *supra* note 56 at 723. (citing Norton, *supra* note 13.)

rejects the anticlassification principle, and in so doing rejects a system that views causes of disparities in mortality, wealth, wages, incarceration, evictions, foster care, and countless others are simply beyond the purview of the law;[446] rejects the choice to be intentionally blind to our society's deep embedded issues with discrimination, oppression, and trauma;[447] and rejects ignoring the long history of American caste systems that has often paved with good intentions as well as insidious.[448]

The reality of cybernetic system discrimination compels the reframing of antidiscrimination law into a law built upon an antisubordination lens and enforced through strict liability. As discussed above in Section IV, courts currently understand disparate treatment discrimination or discriminatory intent as occurring when a human (i) intentionally caused discrimination against someone based on their protected class (ii) by rationally and invidiously considering their protected class, (iii) at the moment of the decision, (iv) while in complete control of their decision-making process. That same section showed that neither the current framework of disparate treatment, nor disparate impact are effective in our cybernetic world.

Therefore, I propose a new understanding of intentional discrimination or disparate treatment as occurring when someone (i) intentionally deployed a system that (ii) caused discrimination against someone based on their protected class. This takes the first and most important element of the intentional discrimination and properly separates it into its two components where *intent* is for identifying who is responsible for the system performance and the *cause* component is the trigger identifying when liability attaches.

To determine who intentionally deployed a system is to look at who was responsible for the decision. It is critical to define "intent" here at the system level. For example, in *Ricci*, the City of New Haven intentionally deployed a testing program. The current antidiscrimination framework in Title VII allows liability only if they intentionally deployed a testing program in order to discriminate. My reformulation restores intent to its clear meaning, separating intent from cause. If they intentionally deployed the testing program, they are responsible for the outcomes and liability will then be based on whether the discriminatory outcomes are sufficient to show that the system caused discrimination based on a protected class.

Given the complex, interdependent nature of cybernetic systems—not to mention the numerous barriers to access, auditing, and testing and evaluation—requiring a person harmed to identify the specific human or machine responsible is too much of a burden. As described below in Sec. V.B. 1, to ensure the burden is appropriately on those in control of these systems, cybernetic discrimination requires a strict liability framework where once a system is found to have caused discrimination, intent is presumed.

Given that liability turns on whether a system is found to have caused discrimination, defining discrimination becomes the turning point in the analysis. "Cause" here is not interested in all the

[446] *Supra* notes 1-7.

[447] Katherine Kirkinis et al., *Racism, racial discrimination, and trauma: a systematic review of the social science literature*, ETHN. HEALTH 1–21 (2018); Arthur P. Brief et al., *Just Doing Business: Modern Racism and Obedience to Authority as Explanations for Employment Discrimination*, 81 ORGAN. BEHAV. HUM. DECIS. PROCESS. 72–97 (2000).

[448] KENDI, *supra* note 39.

specific processes that contributed to the failure, the discrimination. Those concerns are for those responsible for the cybernetic system who want to avoid being liable in the future. Defendants ought to be the ones concerned about the complexities of cybernetic systems, not the plaintiffs who have already shown that the defendant's system harmed them. From the perspective of those harmed, their goal is to not be discriminated against. Therefore, the focus of cause and of antidiscrimination law needs to be on what amounts of disparities between races, genders, sexual orientation, abilities, among others is society willing to tolerate. As shown in Sec. V.B. 2, these are antisubordination questions and exactly what our society needs to address through democratically accountable means. It is the only way we can truly address cybernetic system discrimination.

1. Anti-Discrimination Must Be Enforced Through Strict Liability

Tort law is a common reference point for scholars of antidiscrimination law. First, I will explain why negligence models have failed to enforce anti-discrimination law even in an anti-classification sense. Then, I will show that the reality of cybernetic system discrimination compels anti-discrimination law to be enforced via strict liability, no different than how other cybernetic system failures. From a pure tort law perspective, strict liability is necessary when the consumer cannot adequately enforce the obligations necessary to ensure reasonable quality control.[449] As shown above, there is no way for someone harmed by cybernetic system discrimination, the product of complex, interdependent actions, to possibly enforce the obligations necessary to ensure reasonably antidiscrimination.

a. Negligence is Unable to Address Cybernetic System Discrimination

Negligence requires four elements: an actor owes a duty to conform to a particular standard of care, the actor must breach that duty, the breach must actually and proximately cause legally cognizable harm to the victim. Some scholars have argued that Title VII's anti-discrimination framework is already a form of negligence liability.[450] As a result, negligence frameworks have been a popular suggestion for reforming anti-discrimination law, especially in Title VII.[451] However, by studying the progression of scholarship on negligence and Title VII over time, it is clear that scholars are increasingly concerned about the ability to prove the key elements of intent and causation necessary for anti-discrimination law as the realities of cybernetic system discrimination become unavoidable.[452] As will be shown in Sec. V.B. , a standard rule of torts is that once negligence cannot be proven, then strict liability is necessary.

In 1993, David Oppenheimer proposed a negligence theory of employment discrimination, that an employer should be held liable when "the employer fails to take all reasonable steps to prevent discrimination that it knows or should know is occurring, or that it expects or should expect to occur. An employer should also be found liable when it fails to conform its conduct to the

[449] Geistfeld, *supra* note 59 at 1664. Strict liability also compelled from a moral perspective to ensure anti-discrimination law is achieves accountability. Wirts, *supra* note 19 at 849–50.

[450] David Benjamin Oppenheimer, *Negligent Discrimination*, 141 UNIV. PA. LAW REV. 899, 917 (1993).

[451] For a review, *see* Stephanie Bornstein, *Reckless Discrimination*, 105 CALIF. LAW REV. 1055, 1065 (2017).

[452] *Accord*, Selmi, *supra* note 283 at 770. ("As employers became more sophisticated in their tests, and as the cases moved farther away from the era of overt discrimination, even the testing cases began to fail because it became more difficult for courts to interpret the practices as discriminatory.")

statutorily established standard of care by making employment decisions that have a discriminatory effect, without first carefully examining its processes, searching for less discriminatory alternatives, and examining its own motives for evidence of stereotyping."[453] This properly puts the burden on the defendant-employers. However, in 2014, Richard Ford challenged that "intent" and "causation" are difficult to determine, so he proposed "the law should replace the conceptually elusive goal of eliminating discrimination with the more concrete goal of requiring employers ... to meet a duty of care to avoid unnecessarily perpetuating social segregation or hierarchy."[454] In 2017, Stephanie Bornstein also agreed that showing intent was too difficult and so she proposed a theory of "reckless" discrimination to address the systemic causes of employment discrimination. Reckless discrimination under Title VII would address the "employer entity's responsibility for the widespread operation of implicit bias in the workplace" by assigning liability for an "employer entity's failure to act with sufficient care in creating the context and organizational structures within which employment decisions are made."[455] In sum, authors increasingly belief that the employer's duty must be expanded from adequate pre-deployment examination, to a focus on broad issues of social segregation, and ultimately, responsibility for the full context of cybernetic systems.

In 2019, Andrew Selbst concluded much of this analysis by showing three aspects of the significant gap between how general negligence law operates today and the reality of human-AI systems, or here, cybernetic systems, which leaves those harmed by cybernetic systems without a remedy. First, "[i]t is a fundamental tenet of negligence law that one cannot be liable for circumstances beyond what the reasonable person can account for."[456] As shown in this Essay, the complex, interdependent nature of cybernetic systems makes it difficult for human users to fully understand how their decisions create outcomes, undermining their ability to be "responsible" in the typical sense. This barrier to "foreseeability" and therefore, "reasonableness" is a serious issue for negligence. Selbst suggests that requirements for explainability or interpretability could help alleviate these issues but that unforeseeable or unintuitive outcomes are likely the "rule rather than an exception."[457] As discussed above from a technical perspective, (1) it is not clear when scientists and engineers will be able to achieve explainable machines, but (2) it is very clear that explainable machines will not fully solve the difficulty posed by cybernetic systems.[458]

Second, Selbst correctly explains that although negligence is ideally designed to keep up with technological developments, the opacity, context-dependence, and speed of cybernetic system development is leaving negligence behind.[459] While there are proposals like those discussed above, including access, audits, and testing and evaluation which could identify the most blatant errors and establish best practices, they are already far behind and unlikely to catch up. Moreover, negligence law will still require a distinction between blameworthy errors and non-blameworthy errors which remains a technological and normative barrier when complex, interdependent systems

[453] Oppenheimer, *supra* note 451 at 900.

[454] Richard Thompson Ford, *Bias in the Air: Rethinking Employment Discrimination Law*, 66 STANFORD LAW REV. 1381, 1384 (2014).

[455] Bornstein, *supra* note 452 at 1105.

[456] Andrew Selbst, *Negligence and AI's Human Users*, 100 BOSTON UNIV. LAW REV. 1315, 1360 (2019).

[457] *Id.* at 1362.

[458] *Supra* Sec. V.A. b

[459] Selbst, *supra* note 457 at 1364.

fail.[460]

Third, Selbst discusses how statistical facts that evidence errors or discrimination are not cognizable by a negligence regime focused on individual responsibility.[461] Statistical reasoning in a negligence regime can also perniciously defend a cybernetic system that makes everyone generally better off despite exacerbating disparities – suggesting that specific injuries are less blameworthy.[462] As he concludes "Where injuries are caused by statistical realities, a regime of ex post liability based on fault will simply not be well suited to address the harms."[463]

b. The Need for Strict Liability

Having shown that negligence law cannot account for the traditional intent and causation requirements of anti-discrimination law in the context of cybernetic systems, there is only one solution: strict liability. Under strict liability, if a programming error caused vehicle to crash, the plaintiff would not have to identify the specific programming error.[464] Instead, the plaintiff would only need to "prove defective design solely based on the manner in which the operating system misperformed."[465] Therefore, enforcing antidiscrimination with strict liability would mean that a plaintiff attempting to prove discrimination would not have to identify the specific cause of the discrimination and could instead prove discrimination based on the manner in which the system misperformed. This is the only way to ensure some measure of liability in cybernetic systems because identifying the specific cause of discrimination will often range from too difficult for a typical plaintiff to impossible for teams of experts.

This subsection shows first why failure of negligence compels strict liability. Then it will how the challenges justifying strict product liability for product malfunction in the canonical case of soda bottles and the modern case of autonomous vehicles being hacked, analogize directly to anti-discrimination in cybernetic systems. Lastly, it will show how strict liability also provides value from a moral accountability perspective.

To understand the need for strict liability we can look at the paradigmatic example of product malfunction: the exploding soda bottle. Consumers know that "systems of perfect quality control are either prohibitively expensive or simply unattainable. Some soda bottles will inevitably have undetected problems that cause them to explode (just as food will sometimes be contaminated)."[466] So then why does the exploding soda bottle frustrate the consumer's minimum expectations of safe product performance? Well, consumers still expect a design to be reasonably safe even if the quality control is not perfect. This justifies the standard negligence liability tort rule where the consumer must attribute the misperformance to the manufacturer's failure to exercise reasonable

[460] *Id.* at 1369–70.

[461] *Id.* at 1370.

[462] *Id.* at 1373.

[463] *Id.* at 1374.

[464] Geistfeld, *supra* note 59 at 1634.

[465] *Id.* at 1634, 1634 n. 71. ("RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 3 cmt. B (AM. LAW INST. 1998) (explaining that the plaintiff can recover upon proof of product malfunction without having to "specify the type of defect responsible for the product malfunction.")

[466] *Id.* at 1663.

care in quality control.[467] However, this same expectation of reasonable safety and quality control will justify strict liability when the consumer cannot adequately enforce the obligations necessary to ensure reasonable quality control.[468] The measures for safety could be too complex or cannot be independently evaluated with reliable evidence.[469]

In this case, strict liability is used to solve this evidentiary problem for what would otherwise be a negligence liability. As Oliver Wendell Holmes explained, "the safest way to secure care is to throw the risk upon the person who decides what precautions shall be taken."[470] "Strict liability restores the manufacturer's financial incentive to exercise reasonable care by eliminating the evidentiary barriers to recovery that inhere in the negligence standard. Recognizing as much, the ordinary consumer can reasonably expect compensation for the exploding soda bottle because that form of (strict) liability is necessary for adequately enforcing the manufacturer's underlying obligation to adopt reasonably safe systems of quality control."[471]

Strict product liability still requires that injuries were "caused by a *defect* in the product"[472] so in these cases we make a policy choice that we can make the inference of a defect merely from the "manufacture's *presumed* response to the performance in question."[473] In the case of the soda bottle, one can easily presume that the manufacturer would not have sold the bottle had they known it would explode. Again, negligence may seem appropriate here, but strict liability is necessary when "negligence is hard to prove across the entire category of cases. As previously discussed, the cost and complexity of the negligence inquiry into systems of quality control would often enable the manufacturer to avoid liability as a practical matter. Consequently, the undeniable problem with this aspect of the product's performance—established by the manufacturer's or consumer's presumptive response if they had known about the malfunction—is best addressed by subjecting the manufacturer to strict liability."[474]

Geistfeld analogizes these principles from the soda bottle problem to a cybernetic system, arguing that vehicle manufacturers should be subject to strict liability in the event someone is harmed due to a vehicle being hacked. Because it is fair to presume that the manufacturer would redesign the vehicle to address the vulnerability to prevent the hacking, if vehicle is hacked, it "creates an inference of defect—a malfunction—that provides a defensible basis for (strict) liability that obviates the need for a complex negligence analysis of the vehicle's hardware and software systems."[475] A negligence regime would require analysis of numerous complex, interdependent potential causes of hacking, from the hardware in the engine to the external sensor software. "[P]laintiffs would have to prove what reasonable care requires within a technologically complex and evolving environment. This evidentiary burden is comparable to, if not greater than, the burden faced by a consumer trying to prove that a soda manufacturer failed to adopt reasonably

[467] *Id.* at 1664.
[468] *Id.* at 1664.
[469] *Id.* at 1664.
[470] OLIVER WENDELL HOLMES, JR., THE COMMON LAW 117 (1881).
[471] Geistfeld, *supra* note 59 at 1665.
[472] *Id.* at 1667.
[473] *Id.* at 1667 n. 208.
[474] *Id.* at 1667.
[475] *Id.* at 1668.

safe systems of quality control in the case of an exploding bottle."[476] This Essay emphatically supports that conclusion.

Geistfeld concludes that "[d]ue to the safety problems that would be predictably created by an under-enforced rule of negligence liability, the failure of an operating system to perform in its intended manner due to either a computer bug or third-party hacking provides an inference of defect—a product malfunction—that justifies strict liability. The liability would give the manufacturer the necessary financial incentive for ensuring the reasonable reliability of the operating system. This rule of strict liability only channels a limited number of crashes into the tort system and does not approach the rule of absolute liability that courts have uniformly rejected. For the same reasons that apply to crashes caused by programming bugs, the manufacturer will be subject to strict liability for crashes caused by hacking under the malfunction doctrine or its equivalent, the ordinary consumer expectations test."[477]

Geistfeld's cybernetic cybersecurity problem is directly analogous to the cybernetic anti-discrimination problem at issue here. One can summarize Geistfeld's argument as: (1) vehicle designers do not intend for their vehicle operating systems to cause a vehicle to crash; (2) operating systems can fail due to any number of complex, interdependent reasons; (3) forcing a plaintiff harmed by the operating system failure to prove what reasonable care is necessary to prevent the failure of a complex, interdependent operating system would be an impermissible evidentiary burden; (4) therefore, to ensure the reasonable safety of the vehicle for the general consumer, the plaintiff ought to be able to show liability with only proof that operating system caused the vehicle to crash.

By analogy, one can easily say of government, employers, and others that: (1) those making employment, housing, credit, justice, and other decisions do not intend for their systems to cause discrimination; (2) systems can fail due to any number of complex, interdependent reasons; (3) forcing a plaintiff harmed by the system failure to prove what reasonable care is necessary to prevent the failure of a complex, interdependent system would be an impermissible evidentiary burden; (4) therefore, to ensure the reasonable safety of society for the general public, the plaintiff ought to be able to show liability with only proof that system caused the discrimination.

Therefore, the language of intentional discrimination can stay, but the actual proof of liability must be fundamentally changed from a negligence model to a model of strict liability. Governments, employers, etc. must be strictly liable for *intentionally* deploying a system that *caused* the discriminatory outcomes. This change does not upset the holding by the Supreme Court in *Washington v. Davis* that under the Constitution's Equal Protection Clause, "a law or other official act, without regard to whether it reflects a racially discriminatory purpose, [is not] unconstitutional *solely* because it has a racially disproportionate impact." A plaintiff must prove discriminatory motive on the part of the state actor to receive redress under the Constitution, not just discriminatory impact. Here, under strict liability, a plaintiff must still prove discriminatory intent. It is just that because that evidentiary burden for cybernetic systems is impermissibly high, society will make the policy choice to infer discriminatory motive from proof of disproportionate impact caused by an intentionally deployed cybernetic system.

[476] *Id.* at 1668.
[477] *Id.* at 1668–69.

The typical fears of strict liability can be addressed here, too: that strict liability "could generate an unpredictable, systemic form of extensive liability that would undermine market stability"[478] or "hamper innovation."[479] First, it is not clear how much evidence there is for these concerns, especially in the context of discriminating systems. Second, destabilizing the markets and hampering the current pathways of innovation for systems that produce discrimination would be a clear benefit to society. Requiring cybernetic systems to be accessed, audited, tested, and evaluated prior to deployment is a social good – not unlike our requirements for testing and certification for safety critical systems like drugs, cars, airplanes, or nuclear power plants.[480] True innovation and resulting market success should be for those who show openly through access, audits, tests, and evaluations that their systems can contribute meaningfully to society without discrimination. In fact, strict liability would spur the real innovation needed in the development of methods of design, test, and evaluation to avoid or identify discrimination prior to deployment, and methods for continuous auditing and testing post-deployment.[481]

2. Strict Liability (and Our Society) Needs a Theory of Anti-Subordination

The most critical consequence of a strict liability framework is that it rejects colorblindness and anti-classification and demands a theory of anti-subordination. In order to achieve a world where agents are strictly liable for intentionally deploying systems that cause discriminatory outcomes, the biggest political and societal arguments will be over what constitutes discriminatory outcomes: what type of "discriminatory misperformance" triggers antidiscrimination liability? What amount of disparity in arrest, charges, convictions, plea deals, sentencing, and the death penalty is acceptable? What amount of disparity in wages and pay is acceptable? What amount of disparity in interactions with the family regulation system, investigations, charges, foster care, termination of parental rights is acceptable? These are decidedly anti-subordination questions. Ones our society have avoided for far too long. Ones our society must have answers to. Ones that will require identifying and disturbing underlying power dynamics. As Selmi explained after posing similar questions,

> "One's answer to all of these questions is not likely to turn on whether strict scrutiny or a rational basis review is applied—or whether a theory of intent or impact is used. One's answer will depend on how much discrimination he or she sees in the world, how one interprets ambiguous acts that are subject to varying interpretations. To move courts to see more discrimination would take much more than a new theory or label; it would require persuading them that discrimination explains the observed disparities—but this is precisely the kind of judicial discussion we so rarely have experienced."[482]

Coupling strict liability and antisubordination in order to reform antidiscrimination law

[478] *Id.* at 1623.
[479] Selbst, *supra* note 457 at 1322.
[480] Canellas, *supra* note 38.
[481] PORTER ET AL., *supra* note 400.
[482] Selmi, *supra* note 283 at 770.

responds to Selmi's call for "a more expansive concept of intent" while simultaneously addressing his requirement for "a greater societal commitment to remedying racial, gender and other disparities linked to what is often defined as societal discrimination."[483] This is critical because as Selmi has explained, part of the reason why antidiscrimination law has failed as a doctrine is because it is has been divorced from "a broader social movement designed to delineate the many ways in which intentional discrimination—defined so as not to be limited to animus-based discrimination—continues to influence life choices for so many individuals, particularly minorities and women. Without a sense that discrimination was pervasive, it was simply too difficult for courts to see discrimination other than in the obvious."[484] As exemplified by the turning points in the law around sexual harassment and the rights of LGBTQ+ individuals, "[s]eeking to create a different theory of equality solely through a legal doctrine, one that was in tension with our societal commitments and the interests of elites, was a doomed project. And in some ironic sense, the move to the disparate impact theory perhaps allowed the Supreme Court to see less discrimination, and to remain confined to a conception of intentional discrimination that turned on outdated notions of motive and intent."[485] Redefining antidiscrimination law with an antisubordination lens and enforcing through strict liability is the best way forward if we want to address cybernetic system discrimination.

VI. CONCLUSION

America, and more accurately the justices of the Supreme Court, has chosen the anticlassification principle to understand and enforce antidiscrimination; therefore, choosing to be blind to cybernetic system discrimination. It is long-past time for America to make a commitment to antisubordination, "to eliminate status-based inequality due to membership in those classes, not as a matter of procedure, but of substance."[486] Once again, we should follow the words of Justice Sotomayor: "The way to stop discrimination on the basis of race is to speak openly and candidly on the subject of race, and to apply the Constitution with eyes wide open to the unfortunate effects of centuries of racial discrimination."[487]

This Essay has shown that the reality of cybernetic systems compels the antisubordination principle, enforced by strict liability. When we use machines in our decision making we become a part of cybernetic systems with numerous sources of potential failure and discrimination,[488] that these cybernetic systems are inherently complex and interdependent,[489] which our

[483] *Id.* at 707.

[484] *Id.* at 772–73.

[485] *Id.* at 781.

[486] Barocas and Selbst, *supra* note 13 at 723. (citing Norton, *supra* note 13 at 206, 209.

[487] *Schuette v. Coalition to Defend Affirmative Action, Integration and Immigration Rights and Fight for Equality by Any Means Necessary*, 572 U.S. 291, 381 (2014) (Sotomayor, J., dissenting) ("In my colleagues' view, examining the racial impact of legislation only perpetuates racial discrimination. This refusal to accept the stark reality that race matters is regrettable. The way to stop discrimination on the basis of race is to speak openly and candidly on the subject of race, and to apply the Constitution with eyes open to the unfortunate effects of centuries of racial discrimination. As members of the judiciary tasked with intervening to carry out the guarantee of equal protection, we ought not sit back and wish away, rather than confront, the racial inequality that exists in our society. It is this view that works harm, by perpetuating the facile notion that what makes race matter is acknowledging the simple truth that race does matter.")

[488] *Supra* Section **Error! Reference source not found.**

[489] *Supra* Section III.B.

antidiscrimination laws are incapable of understanding[490] and which no amount of popular proposals can fully address.[491] Strict liability is necessary to enforce liability for discrimination because identifying the specific source of discrimination would be too much of an evidentiary burden on the plaintiffs.

To conclude this paper, it is crucial to understand that this cybernetic black hole poses problems for many more aspects of law than antidiscrimination. While this paper focuses explicitly on the question of anti-discrimination, that is but one example of many where the nature of cybernetic systems is incompatible with our legal system. The law is littered with cybernetic black holes. Just looking at "intent," intent is not unique to antidiscrimination law. Understanding a law enforcement officer's intent is critical to the good faith[492] and deadly force[493] exceptions to the Fourth Amendment protection against unwarranted search and seizure and the public safety exception to the Fifth Amendment requirements of the Miranda warning.[494] Separately, given the interdependence and complexity of cybernetic systems, who or what does the Constitution require be available for confrontation by a criminal defendant?[495] Or, what procedures and information are necessary to ensure that scientific evidence is reliable under *Daubert* and *Frye*[496] or that an individual will receive procedural due process?[497]

In the words of Norbert Weiner, the father of cybernetics: "Whether we entrust our decisions to machines of metal, or to those machines of flesh and blood which are bureaus and vast laboratories and armies and corporations, we shall never receive the right answers to our questions unless we ask the right questions. ... The hour is very late, and the choice of good and evil knocks at our door."[498]

[490] *Supra* Section IV.

[491] *Supra* Section V.A.

[492] *United States v. Leon*, 468 U.S. 897 (1984); *Davis v. United States*, 564 U.S. 229 (2011). Good faith depends, in part, upon whether law enforcement is the source of the error for the violation, *Arizona v. Evans*, 514 U.S. 1 (1995), and the level of negligence, *Herring v. United States*, 555 U.S. 135 (2009).

[493] *Graham v. Connor*, 490 U.S. 386 (1989)

[494] *New York v. Quarles*, 467 U.S. 649 (1984)

[495] *Williams v. Illinois*, 567 U.S. 50 (2012)

[496] *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923); *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993); *General Electric Co. v. Joiner*, 522 U.S. 136 (1997); *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999); Canellas, *supra* note 38. *See Id.*

[497] *Mathews v. Eldridge*, 424 U.S. 319 (1976); Danielle Keats Citron, *Technological Due Process*, 85 WASH. UNIV. LAW REV. 1249 (2008). (explaining how *Mathews v. Eldridge* cost-benefit analysis won't allow technological due process because of the cost of expert analysis)

[498] NORBERT WIENER, THE HUMAN USE OF HUMAN BEINGS: CYBERNETICS AND SOCIETY 185–86 (1989 ed. 1950).