UC DAVIS LAW REVIEW ONLINE

VOL. 52 APRIL 2019

Debating Algorithmic Fairness*

Melissa Hamilton[**]

TABLE OF CONTENTS

[*] To view the entire article as the author intended, please see the annotated version at https://hyp.is/go?url=https://lawreview.law.ucdavis.edu/online/vol52/52-online-Hamilton.pdf&q=user:qdr@hypothes.is. The Project will incorporate the novel methodology of annotation for transparent inquiry. *Annotations* are external to the main text, designed to be supplemental attributions that allow the reader to better assess the quality and rigor of the research and conclusions. An annotation commonly will include some combination of: (a) an analytical note (e.g., to provide further description or explanation of the process and choices the researcher made, to comment on any assumptions the researcher has made, or delineate other contextual underpinnings); (b) expanded source quotations; and/or (c) further citations or discourse as evidence of claims made. Also, an annotation can take advantage of technological developments by embedding hyperlinks to web-based sources that are more readily available for readers to access. Access to a stand-alone copy of the annotations and underlying data are available here: https://doi.org/10.5064/F6JOQXNF.

[**] Copyright © 2019 Melissa Hamilton. Senior Lecturer of Law & Criminal Justice, University of Surrey; J.D., The University of Texas School of Law; Ph.D. (criminology), The University of Texas at Austin.

INTRODUCTION

Automated risk assessment is the trendy model in the criminal justice system.[1] Proponents view risk assessment as an objective and reasonable way to reduce mass incarceration without sacrificing public safety.[2] Professor Christopher Slobogin refers to the practice as "preventive justice."[3] Officials thus are becoming heavily invested in risk assessment tools — along with their reliance upon big data and algorithmic processing — to inform decisions on managing offenders according to their risk profiles.[4]

Nonetheless, a public debate on the topic emerged when the investigative journalist group ProPublica recently proclaimed that a popular risk tool called COMPAS was racially biased.[5] Analyzing COMPAS's performance on a large dataset, ProPublica found that blacks who did not reoffend "are almost twice as likely as whites to be labeled a higher risk," while whites who reoffended "are much more likely than blacks to be labeled lower risk."[6] In statistical terms, these results mean that the tool produced higher false positive rates for blacks and higher false negative rates for whites.

Prominent news sites highlighted ProPublica's message that this proved yet again an area in which criminal justice outcomes were racist, even despite using a mathematical algorithm.[7] COMPAS's

[1] J. Stephen Wormith, *Automated Offender Risk Assessment: The Next Generation or a Black Hole?*, 16 CRIMINOLOGY & PUB. POL'Y 281, 281 (2017).

[2] *See generally* Brandon L. Garrett, *Evidence-Informed Criminal Justice*, 86 GEO. WASH. L. REV. 1490 (2019) (providing a background on criminal justice advancements, and analyzing, as well as advocating, empirically based approaches to adopting policy).

[3] Christopher Slobogin, *Preventive Justice: A Paradigm in Need of Testing*, BEHAV. SCI. L. 1, 1 (2018).

[4] *See* Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L. J. 1147, 1149 (2017) ("Today's algorithms are digital 'robots' that possess effectively autonomous abilities to adopt and learn"; a "type of artificial intelligence."). *See generally* Nathan James, Cong. Res. Serv., *Risk and Needs Assessment in the Criminal Justice System* (Oct. 13, 2015), https://digital.library.unt.edu/ark:/67531/metadc795663/m1/1/high_res_d/R44087_2015Oct13.pdf.

[5] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[6] *Id.*

[7] *E.g.*, Li Zhou, *Is Your Software Racist?*, POLITICO (Feb. 7, 2018, 5:05 AM), https://www.politico.com/agenda/story/2018/02/07/algorithmic-bias-software-recommendations-000631 (discussing, in part, how racially based bad data creates racially based bad algorithms); Ed Yong, *A Popular Algorithm Is No Better at Predicting Crime than Random People*, ATLANTIC (Jan. 18, 2018), https://www.theatlantic.com/technology/archive/

corporate owner, Northpointe, denied the allegations, stating that its reanalysis of the dataset ProPublica used led to the contrary conclusion: the tool was unbiased as blacks and whites had similar positive predictive values for recidivism.[8]

This debate is credited with sparking a movement to promote algorithmic fairness.[9] As a recent illustration, a consortium of over 100 legal organizations, government watchdog groups, and minority rights associations (e.g., ACLU, NAACP, and Electronic Frontier Foundation) signed onto "A Shared Statement of Civil Rights Concerns." In doing so, they were expressing unease that algorithmic tools may create disparate impact based on race or other protected characteristics.[10]

The ProPublica/Northpointe racial bias debate, and the broader issue of algorithmic fairness, present significant dilemmas for criminal justice officials, legal practitioners, data scientists, and policymakers.

> There remains significant opportunity to influence and manage the development of computer technology, to ensure that ethics and law are part of the curriculum of software developers and analysts, and to regulate as necessary. However, the development of big data, computer-assisted decision-making, and e-regulation present serious challenges

2018/01/equivant-compas-algorithm/550646/ (discussing how in one such study from Dartmouth College, "COMPAS [was] no better at predicting an individual's risk of recidivism than random volunteers recruited from the internet"); Max Ehrenfreund, *The Machines that Could Rid Courtrooms of Racism*, WASH. POST (Aug. 18, 2016), https://www.washingtonpost.com/news/wonk/wp/2016/08/18/why-a-computer-program-that-judges-rely-on-around-the-country-was-accused-of-racism/?noredirect=on&utm_term=.ce854f237cfe (discussing the paradox of how "[s]ystemic racial injustices can be reflected in software that holds the promise of greater equality"); NPR, *The Hidden Discrimination in Criminal Risk-Assessment Scores* (May 24, 2016), https://www.npr.org/2016/05/24/479349654/the-hidden-discrimination-in-criminal-risk-assessment-scores (discussing, via the program All Things Considered, how "[c]ourtrooms across the country are increasingly using a defendant's "risk assessment score" to help make decisions about bond, parole and sentencing").

[8] William Dieterich et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, NORTHPOINTE INC. RES. DEP'T (July 8, 2016), http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

[9] Chelsea Barabas et al., *Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment*, 81 PROC. MACHINE LEARNING RES. 1, 1 (2018), proceedings.mlr.press/v81/barabas18a/barabas18a.pdf; Thomas Miconi, *The Impossibility of "Fairness": A Generalized Impossibility Result for Decisions* 3 (Sep. 11, 2017), https://pdfs.semanticscholar.org/d883/b155d1ce19672cdf49795ea1a63acc923ad5.pdf.

[10] African American Ministers in Action et al., *The Use of Pretrial "Risk Assessment" Instruments: A Shared Statement of Civil Rights Concerns* 9-10 (2018), http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf.

> to the rule of law, equality, and natural justice, and the poor understanding and transparency of software development means that this requires serious attention from those who research, teach, and practice law.[11]

This paper addresses these concerns with a mixed-methods empirical project to inform legal practitioners and data scientists on some of these difficult challenges. The Article proceeds as follows. Part I summarizes the background for the study. A focal point is to investigate how ProPublica and Northpointe came to such contrasting conclusions, despite studying the use of a single tool with the same dataset.

Part II introduces the methodology. The qualitative component engages a critical discourse analytic approach to adjudicate the debate. The discursive argumentation here involves two powerful groups. ProPublica has access to public influence as a media outlet, while Northpointe is a machine-learning company which lends itself to scientific cachet. Each may be strategically attempting to control the wider understandings of the utility and parity of algorithmic risk assessment. Part II also describes the dataset that is mined in order to offer a quantitative supplement. ProPublica and Northpointe have utilized only a few of the algorithmic fairness measures that data scientists and legal experts now advance. Thus, the statistical offering to this debate includes computing several of these alternative equations to supply additional evidence as to the potential for racial bias in the COMPAS risk tool. To do this, we use the same dataset again so that any differentials are not due to discrepancies in sample/population profiles.

Part III relays the results of the qualitative analysis, followed by the quantitative supplement. The information therein unpacks the ProPublica/Northpointe dispute in terms of its contrasting verdicts. Several algorithmic fairness measures the parties ignore are identified and computed. Part III acts as a third-party audit and a window into

[11] Rónán Kennedy, *Algorithms and the Rule of Law*, 17 LEG. INFO. MGMT. 170, 172 (2017).

> While far from a panacea, data mining can and should be part of a panoply of strategies for combating discrimination . . . and for promoting fair treatment and equality. Ideally, institutions can find ways to use data mining to generate new knowledge and improve decision making that serves the interests of both decision makers and protected classes.

Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 732 (2016).

transparency regarding a popular algorithmic risk tool. Independent conclusions are rendered as to the potential for disparate impact with respect to race.

I. STUDY BACKGROUND

Risk assessment in criminal justice entails predicting an individual's potential for recidivism in the future.[12] Predictions have long contributed to criminal justice decision-making as serving the legitimate goal of protecting the public from those who have been identified as offenders.[13] Historically, risk predictions typically relied upon instinct or the limited personal experience of the official responsible for making the relevant decision.[14] A wave of more empirically informed risk assessment tools has emerged, aided by advances in behavioral sciences, the availability of big data, and improvements in statistical modeling.

A. The New Wave Risk Assessment in Criminal Justice

The "evidence-based practices movement" is the now popular term to describe the turn to behavioral sciences data to improve risk-based classifications.[15] Scientific studies targeting recidivism outcomes are benefiting from the compilation of large datasets (i.e., big data) of discharged offenders. Researchers track the offenders post-release, observe recidivism rates, and then statistically test which factors correlate with recidivism.[16] Risk assessment tool developers use computer modeling to combine factors of sufficiently high correlation and weight them accordingly using increasingly complex algorithms.[17]

[12] Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231, 232 (2015).

[13] Jordan M. Hyatt et al., *Reform in Motion: The Promise and Perils of Incorporating Risk Assessments and Cost-Benefit Analysis into Pennsylvania Sentencing*, 49 DUQ. L. REV. 707, 724-25 (2011).

[14] Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 556 (2015).

[15] Faye S. Taxman, *The Partially Clothed Emperor: Evidence-Based Practices*, 34 J. CONTEMP. CRIM. JUST. 97, 97-98 (2018).

[16] Kelly Hannah-Moffat, *Algorithmic Risk Governance: Big Data Analytics, Race and Information Activism in Criminal Justice Debates*, THEORETICAL CRIMINOLOGY (March 22, 2018), https://doi.org/10.1177/1362480618763582.

[17] An algorithm refers to "computation procedures (which can be more or less complex) drawing on some type of digital data ('big' or not) that provide some kind of quantitative output (be it a single score or multiple metrics) through a software program." Angéle Christin, *Algorithms in Practice: Comparing Web Journalism and*

Broadly speaking, "[d]ata-driven algorithmic decision making may enhance overall government efficiency and public service delivery, by optimizing [bureaucratic] processes, providing real-time feedback and predicting outcomes."[18] With such a statistical tool in hand, criminal justice officials can more consistently input relevant data and receive software-produced risk classifications.[19]

The utility of risk instruments has attracted energetic support from reputable policy centers and been the subject of various legislative policy proposals as a driver of criminal justice reform.[20] News headlines and academic literature have also been expounding upon the benefits generated by the government's use of big data to predict the future risk posed by individuals.[21] Algorithmic risk assessment tools offer the ability to reduce mass incarceration by diverting low-risk defendants from prison, while targeting greater supervision and services to those at higher risk.[22]

Many parties presume that algorithmic risk assessment tools developed on big data epitomize transparent, consistent, and logical methods for classifying offenders.[23] The mathematical character of risk assessment suggests the ability to quantify the future and transport it into the present.[24] Evidence-based practices thereby present a welcome displacement of human instinct.[25]

Criminal Justice, BIG DATA & SOC'Y 1, 2 (July–Dec. 2017), http://journals.sagepub.com/doi/pdf/10.1177/2053951717718855.

[18] Bruno Lepri et al., *Fair, Transparent, and Accountable Algorithmic Decision-Making Processes*, 31 PHIL. & TECH. 611, 611-12 (2018), www.nuriaoliver.com/papers/Philosophy_and_Technology_final.pdf.

[19] Wormith, *supra* note 1, at 285.

[20] *See* Erin Collins, *Punishing Risk*, 107 GEO. L.J. 57, 57 (2018). *See generally* James, *supra* note 4.

[21] *E.g.*, Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PENN. L. REV. 327, 407 (2015); Crysta Jentile & Michelle Lawrence, *How Government Use of Big Data Can Harm Communities*, FORD FOUND. (Aug. 30, 2016), https://www.fordfoundation.org/ideas/equals-change-blog/posts/how-government-use-of-big-data-can-harm-communities/; Sony Kassam, *Legality of Using Predictive Data to Determine Sentences Challenged in Wisconsin Supreme Court Case*, A.B.A. J. (June 27, 2016, 1:07 PM), http://www.abajournal.com/news/article/legality_of_using_predictive_data_to_determine_sentences_challenged_in_wisc.

[22] Sarah L. Desmarais et al., *Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings*, 13 PSYCHOL. SCI. 206, 206-07 (2016).

[23] Hyatt et al., *supra* note 13, at 725.

[24] M. Roffey & S.Z. Kaliski, *To Predict or Not to Predict — That Is the Question*, 15 AFR. J. PSYCHIATRY 227, 227 (2012).

[25] Alfred Blumstein, *Some Perspectives on Quantitative Criminology Pre-JQC: and Then Some*, 26 J. QUANTITATIVE CRIMINOLOGY 549, 554 (2010).

Nonetheless, concerns have arisen whether algorithmic tools are as fair as expected, particularly concerning the potential for disparate impact on protected groups.[26] An important instigator is a report released in 2016 that identified a widely used software risk tool as producing racist predictions.

B. Competing Claims Concerning the Racist Algorithm

Investigative news journalists with the nonprofit ProPublica reported on statistical analyses the group had conducted involving a real dataset and a popular risk tool named COMPAS — the acronym for Correctional Offender Management Profiling for Alternative Sanctions. ProPublica investigators obtained (through Freedom of Information Act requests) the data on over 7,000 arrestees who were scored on COMPAS in a pretrial setting in a southern county of Florida.[27] These scores had previously been provided to judges as evidence to consider when ruling on pretrial release for individual arrestees.[28] ProPublica concluded COMPAS was racist in that its algorithm produced a much higher false positive rate for blacks than whites (45% versus 24%, respectively), meaning that it overestimated high risk for blacks.[29]

COMPAS's corporate owner, Northpointe, quickly rejected such characterizations.[30] After running their own statistical analyses on the same dataset ProPublica had compiled, Northpointe statisticians asserted that their results demonstrated COMPAS outcomes achieved predictive parity for blacks and whites.[31] More specifically, Northpointe reported that black defendants who were predicted to

[26] Sandra Mayson, *Bias In, Bias Out*, 128 YALE L.J. (forthcoming 2019); Christin, *supra* note 17; Osonde Osaba & William Welser IV, RAND CORP., *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence* 19 (2017), https://www.rand.org/pubs/research_reports/RR1744.html.

[27] Angwin et al., *supra* note 5.

[28] *See* Malcolm M. Feeley, *How to Think About Criminal Court Reform*, 98 B.U. L. REV. 673, 683 (describing big data-based risk assessment in a pretrial setting an example of the new wave of criminal court reform).

[29] Angwin et al., *supra* note 5.

[30] Northpointe rebranded with the trade name equivant (lower case intended) in January 2017. Press Release, equivant, Courtview, Constellation & Northpointe Re-brand to equivant (2017), http://www.equivant.com/blog/we-have-rebranded-to-equivant.

[31] William Dieterich et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity* 2-3 (July 8, 2016), http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

recidivate did reoffend at a "slightly" higher rate than whites (63% versus 59%, respectively).[32]

It turns out that the dispute is founded upon contrasting measures of algorithmic fairness. ProPublica touted the false positive rate. Northpointe instead preferred the alternative measure called the positive predictive value.[33] Still, the potential that a popular risk tool used in a criminal justice setting to inform release decisions is racist (or not) attracted much attention.[34] And, because of the contrary conclusions on that question, academics, practitioners, and the media were evidently confused.[35] One of the purposes of this paper is to offer appropriate explanations to relieve such misunderstandings. But for now, perhaps a brief summary will suffice. The rates of reoffending between groups varied significantly, such that the base rate of rearrests for blacks was much higher than for whites in the underlying dataset. The obstacle is that when base rates between groups differ, the algorithm cannot achieve equal false positive rates and equal positive predictive values at the same time because only the latter statistic is heavily influenced by base rate differentials.[36]

The subject of group-based differences in prediction is even more complicated than the few conflicting statistical measures that

[32] *Id.* at 11, 20.

[33] Tafari Mbadiwe, *Algorithmic Injustice*, NEW ATLANTIC 1, 18 (Winter 2018), https://www.thenewatlantis.com/publications/algorithmic-injustice; *see also* Alexandria Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 155 (2017).

[34] *E.g.*, Catherine Matacic, *Are Algorithms Good Judges?: People are as Good as Machines in Predicting Rearrest*, 359 SCI. 263, 263 (2018); Carole Piovesan & Vivian Ntiri, *Adjudication by Algorithm: The Risks and Benefits of Artificial Intelligence in Judicial Decision-Making*, ADVOCATES J. 42-43 (Spring 2018); Adam Liptak, *Sent to Prison by a Software Program's Secret Algorithms*, N.Y. TIMES (May 1, 2017), https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html.

[35] *E.g.*, Taylor R. Moore, Ctr. Democracy & Tech., *Trade Secrets and Algorithms as Barriers to Social Justice* (Aug. 2017), https://cdt.org/. . .//files/2017/08/2017-07-31-Trade-Secret-Algorithms-as-Barriers-to-Social-Justice.pdf; Jason Tashea, *Risk-Assessment Algorithms Challenged in Bail, Sentencing and Parole Decisions*, ABA J. (Mar. 1, 2017 01:30 CST), http://www.abajournal.com/magazine/article/algorithm_bail_sentencing_parole/; Joshua Brustein, *This Guy Trains Computers to Find Future Criminals*, BLOOMBERG (July 18, 2016), https://www.bloomberg.com/features/2016-richard-berk-future-crime/; Ryan O'Hare, *Is Software Used by Police to Identify Suspects Racist?*, DAILY MAIL (May 24, 2015 07:51 EDT), https://www.dailymail.co.uk/sciencetech/article-3606478/Is-software-used-police-identify-suspects-racist-Algorithm-used-predict-likelihood-reoffending-biased-against-black-people-investigation-claims.html.

[36] Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, SOC. METHODS & RES. 18 (forthcoming 2019), https://arxiv.org/pdf/1703.09207.

ProPublica and Northpointe highlighted. Additional computations for algorithmic equity now exist within the legal and scientific literatures, albeit some of them being mutually exclusive as well.[37] Several of these alternative measures of algorithmic fairness will be discussed and quantified herein.

The next Part sets up the analytical plan underlying the mixed-methods research reported herein. The overall intent is to tease out how ProPublica and Northpointe came to their competing notions of fairness, examine justifications for their choices, and then supplement their narrow perspectives on algorithmic equity.

II. Methodology

This interdisciplinary study of algorithmic risk prediction combines qualitative and quantitative methods. The analytical plan draws upon multiple sources of data, both textual and statistical, that complement each other for a more integrated approach. The qualitative component embraces critical discourse analysis as its methodological framework.

A. *Critical Discourse Analysis*

Discourse is communicative, yet it should not be taken at face value. Discourse analysts understand that people's communication can be strategic in attempting to exercise control over mutual understandings of the issue at hand.[38] Discourse analysis is thereby interested in many things: why text is framed as it is; why certain words are used and in what order; what the specific text implies about broader discourses; and how in the larger scheme of things the discourse reflects and conveys information about social structures and power.[39] Discourse analysis is particularly suited, therefore, to communications research into argumentation.[40]

Critical discourse analysis ("CDA") extends discourse analysis: "Rather than merely *describe* discourse structures, it tries to *explain* them in terms of properties of social interaction and especially social

[37] Miconi, *supra* note 9, at 2.

[38] Karen Tracy, *Discourse Analysis in Communication*, *in* 4 THE HANDBOOK OF DISCOURSE ANALYSIS 725, 731 (Deborah Schiffrin et al. eds., 2008).

[39] Christina Schäffner, *Discourse Analysis*, *in* 4 HANDBOOK OF TRANSLATION STUDIES 47, 48 (Yves Gambier et al. eds., 4th ed. 2013).

[40] Tracy, *supra* note 38, at 732 ("[T]o identify how relatively stable aspects of meaning are acted upon by the shaping and changing power of context . . . . [using] forecasting principles which communicators use to make decisions about what to say next is identified.").

structure."[41] CDA is acutely engaged in how discourse is deployed to maximize imbalances in power.[42] Thus, CDA represents "discourse analytical research that primarily studies the way social-power abuse and inequality are enacted, reproduced, legitimated, and resisted by text and talk in the social and political context."[43]

CDA can tease out how dominant players manipulate the text and context of public discourse with the aim of "controlling the intentions, plans, knowledge, opinions, attitudes, and ideologies — as well as their consequent actions — of recipients."[44] The use of epistemic or ideological manipulation often instrumentally serves the speaker's own interests.[45] The powerful may dictate the dialogue by exploiting their guise of authority and credibility, particularly when the audience does not have the ability or knowledge to challenge it.[46] Discursive-controlling power strategies at times may simply be aimed at tempering the voices of others.[47]

When a powerful group is relying on perceptions of its authority to control, repeated exercises in cultivating confidence in the group's legitimacy may appear necessary.[48] Recognized types of legitimizing accounts include authorization, moral evaluation, and rationalization.[49]

On a microanalytical front, discourse control practices can employ tactical choices with specific lexical choices, syntactic structures, rhetorical devices, or narrative arrangements.[50] For example, when a group perceives that another party's message is challenging the group's authority, useful lexical expressions can entail words or labels that are

[41] Teun A. van Dijk, *Critical Discourse Analysis*, *in* THE HANDBOOK OF DISCOURSE ANALYSIS 466, 467 (Deborah Tannen et al. eds. 2d ed., 2015) [hereinafter *Critical Discourse Analysis*].

[42] Teun A. van Dijk, *Critical Discourse Studies: A Sociocognitive Approach*, *in* METHODS OF CRITICAL DISCOURSE STUDIES 62, 71 (Ruth Wodak & Michael Meyer eds., 3d ed. 2015) [hereinafter *Critical Discourse Studies*].

[43] van Dijk, *Critical Discourse Analysis*, *supra* note 41, at 466.

[44] *Id.* at 472.

[45] *Id.* at 470, 473.

[46] *Id.* at 470, 473.

[47] *Id.* at 470-72.

[48] Theo van Leeuwen, *The Discursive Construction of Legitimation*, *in* DISCOURSE AND PRACTICE: NEW TOOLS FOR CRITICAL ANALYSIS 105, 105 (Nikolas Coupland & Adam Jaworsk eds., 2008) [hereinafter *Discursive Construction*] (citing Max Weber).

[49] *Id.* at 106-07.

[50] van Dijk, *Critical Discourse Analysis*, *supra* note 41, at 471-73.

judgmental in nature and meant to restrict how the other party's message is perceived, both internally and externally.[51]

Considering the foregoing, CDA researchers are keen to test the claims of powerful groups engaged in discourse from a validity perspective. More specifically, validity claims identify four attributes: legitimacy, truthfulness, sincerity, and comprehensibility.[52] A key measure of legitimacy arises if the speaker properly recognizes and considers different perspectives. In terms of truthfulness, a particular strength of the CDA method is its evaluative component. The researcher may uncover misrepresentations which appear to protect the speaker's own interests, while also maintaining the power imbalance.[53] One approach for establishing that a discursive turn constitutes a misrepresentation involves exposing internal inconsistencies or contradictory positions.[54] In sum, as a prominent qualitative scholar observes, "CDA is discourse analysis *with an attitude*."[55]

The CDA results will then be supplemented by algorithmic fairness calculations from a live dataset. Next is a summary of this dataset and the measures underlying the quantitative aspect of the study presented herein.

B. Dataset and Measures

The primary dataset for the quantitative piece includes individuals arrested in Broward County, Florida and scored on the COMPAS general recidivism risk scale after their arrests in 2013 and 2014.[56] Notably, Broward County is among the top twenty largest American counties by population,[57] thus improving the potential for a large and

[51] *Id.* at 473.

[52] Jeffrey D. Wall et al., *Critical Discourse Analysis as a Review Methodology: An Empirical Example*, 37 COMM. ASS'N INFO. SYS. 257, 261 (Sept. 2015), http://elibrary.aisnet.org/Default.aspx?url=https://aisel.aisnet.org/cgi/viewcontent.cgi?article=3876&context=cais.

[53] Theo van Leeuwen, *Moral Evaluation in Critical Discourse Analysis*, 15 CRITICAL DISCOURSE STUD. 140, 144 (2018) [hereinafter *Moral Evaluation*].

[54] *Id.* at 144.

[55] van Dijk, *Critical Discourse Analysis*, *supra* note 41, at 466.

[56] *See* Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), http://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. ProPublica has generously made the data available for other researchers to access. *Data Analysis for 'Machine Bias,'* GITHUB (June 12, 2017), http://github.com/propublica/compas-analysis.

[57] Matt Rosenberg, *Largest Counties by Population*, THOUGHTCO. (Feb. 12, 2018), https://www.thoughtco.com/largest-counties-by-population-1435134.

diverse sample set. The pretrial services division of the Broward County Sheriff's Office has been using COMPAS since 2008 to inform judicial determinations concerning pretrial release.[58]

COMPAS is a software application widely used across correctional institutions and offers a general recidivism risk scale.[59] The COMPAS algorithm produces outcomes as decile scores of 1–10 with higher deciles representing greater predicted risk. COMPAS then subdivides decile scores into three, ordinal risk bins: low (deciles 1–4), medium (deciles 5–7), and high (deciles 8–10).[60]

This study uses a subset of the data in which individuals were scored on the general recidivism risk tool within thirty days of their arrests and for whom two years of follow-up after release were available, *n* = 6,172. As the focus is on comparing blacks and whites, individuals who were not in those racial groupings were omitted, leaving a sample size of *n* = 5,278.

III. RESULTS

This is a mixed-methods study to better elaborate on the issues raised concerning test bias and disparate impact by race. The qualitative results from a critical discourse analysis of the Northpointe/ProPublica debate are offered first and then the quantitative results follow. This ordering is appropriate considering one of the critiques in the qualitative portion is that the main discourses fail to address several of the popular algorithmic risk fairness definitions. Quantitative results that derive from data analyses and reported herein fill that gap. The statistical offerings thereby supplement the qualitative discourse evaluation, provide additional information on the abilities of COMPAS, and contextualize further why various algorithmic fairness definitions can provide what appear to be inconsistent results.

To begin, the qualitative component relies upon critical discourse analysis methods, as summarized earlier.

[58] THOMAS BLOMBERG ET AL., VALIDATION OF THE COMPAS RISK ASSESSMENT CLASSIFICATION INSTRUMENT 15-16 (2010).

[59] EQUIVANT, PRACTITIONER'S GUIDE TO COMPAS CORE 4 (Dec. 19, 2017), http://equivant.volarisgroup.com/assets/img/content/Practitioners_Guide_COMPASCore_121917.pdf.

[60] *Id.* at 8.

A. *Qualitative Results*

The CDA results are divided among three general categories. The first provides support for how the ProPublica/Northpointe debate sets itself up as conflict speech. In consideration of space limitations, the remainder of the report is circumspect in two ways. Northpointe's perspective is the main concern considering its inherent conflict of interest due to its ownership of COMPAS. Then the discussion and quantitative results hone on the COMPAS general recidivism risk scale rather than try to also cover the alternative COMPAS violent recidivism scale.

Then, the second component of the qualitative study reviews the main strategies underlying Northpointe's evident attempts to control the message about COMPAS and racial bias. The third category observes the discursive strategies in terms of what otherwise are important and relevant issues and subjects, yet which Northpointe expediently neglect. These omissions will then, in large measure, be remedied in the quantitative results section.

1. The Setup for the Discursive Conflict

In news stories and scholarly articles, the use of headlines or titles can present as an inaugural signal of a structural scheme meant to persuade the reader towards the author's predetermined conclusion.[61] It is relatively easy to establish that ProPublica and Northpointe at the outset are engaging in conflict speech in which each is attempting through its discourses to exercise control over public understandings about bias in the COMPAS algorithm.[62] ProPublica begins the exchange, crafting a title to its initial report evidently designed to grab attention: "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks."[63] Here, ProPublica's initial publication engages the socially powerful narrative of racial discrimination.[64] Northpointe's quick response was to post a document on its corporate website, utilizing a title to explicitly stake its position, though in less provocative terms: "COMPAS Risk Scales:

[61] *See* van Dijk, *Critical Discourse Studies*, *supra* note 42, at 72.

[62] Communications scholars tend to define conflict in terms of incompatibility of goals and values, an expression of struggle between interested parties, and some interdependence between them. Linda L. Putman, *Definitions and Approaches to Conflict and Communication*, *in* THE SAGE HANDBOOK OF CONFLICT COMMUNICATION 1, 5 (John G. Oetzel & Stella Ting-Toomey eds., 2006).

[63] Angwin et al., *supra* note 5.

[64] *See* van Dijk, *Critical Discourse Analysis*, *supra* note 41, at 468.

Demonstrating Accuracy Equity and Predictive Parity."[65] For the sake of limited space, the qualitative analysis herein refers to the two documents just mentioned, unless otherwise specified.

This critical discourse analysis is rather unique in the sense that there is not *one* voice in the conflict holding dominating authority.[66] Instead, both parties enjoy a consequential level of group power, though the sources of power are distinct. ProPublica carries the power of the news media to direct discourse among the general public.[67] Northpointe's guise of science as an analytical software company holds a different power, one that may be able to control scientific discourse.[68] Northpointe, though, evidently understands the power of the media and the stakes at hand. On the first page of its response, Northpointe clearly references ProPublica's conclusions concerning racial bias which, it decries, "were repeated subsequently in interviews and in articles in the national media."[69]

Northpointe, despite its descriptively sober paper title, adopts lexical choices bearing judgmental labels and narrative structures in an evident attempt to counter ProPublica's advances, undermine the media group's scientific abilities, and reauthenticate the COMPAS tool's credibility. In the introduction of its response, Northpointe makes several proclamations. The first is to quickly and summarily establish its rival's scientific failings: "Our review leads us to believe that ProPublica made several statistical and technical errors such as . . . wrongly defined classification terms and measures of discrimination, and the incorrect interpretation and use of model errors."[70] Notice that Northpointe's attempt to retrieve authority here includes denouncing ProPublica's efforts by using righteous terminology in variations of the words "error," "wrong," and

[65] Dieterich et al., *supra* note 8.

[66] van Dijk, *Critical Discourse Analysis*, *supra* note 41, at 470 ("[M]embers of more powerful social groups and institutions, and especially their leaders (the *symbolic elites*), have more or less exclusive access to, and control over, one or more types of public discourse.") (citations omitted).

[67] ProPublica's involvement in a CDA paper is supported by its mission statement: "To expose abuses of power and betrayals of the public trust by government, business, and other institutions, using the moral force of investigative journalism to spur reform through the sustained spotlighting of wrongdoing." *About Us*, PROPUBLICA, https://www.propublica.org/about (last visited Mar. 2, 2019).

[68] *See* van Dijk, *Critical Discourse Analysis*, *supra* note 41, at 469 (referencing the media and science as powerful resources with control over specific forms of discourse).

[69] Dieterich et al., *supra* note 8, at 1.

[70] *Id.*

"incorrect" — and all in a single sentence. It is evident that Northpointe's document is attempting to mediate ProPublica's text and the veracity of its message in the public forum.[71]

Then Northpointe seeks, through the guise of its own authority, to manipulate the audience. The first page of its response states that "[w]hen the correct classification statistics are used, the data do not substantiate the ProPublica claim of racial bias towards blacks."[72] Further, Northpointe promotes the legitimacy of its own moral evaluation, offering that "[t]he proper interpretation of the results in the samples used by ProPublica demonstrates that the General Recidivism Risk Scale . . . [is] equally accurate for blacks and whites."[73] Hence, its discourse discounts ProPublica's statistical finding of racial bias by instead recharacterizing it as merely a "claim." Then it quickly confirms the Northpointe team's purportedly better grasp at statistical skills by using such legitimizing terms as having used the "correct" statistics and making a "proper" interpretation.

Thereby, the conflict structure of the discourse is set. The argumentative positioning acts as a referential foundation for Northpointe's discursive strategies identified next.

2. Discursive Strategies

In conflict discourse terms, a communicative style may strategically assume a rhetorical approach, even via exaggeration, to persuade by engaging "discourse through which a speaker presents an intact monologue supporting a disputable position."[74] Northpointe's self-interest as the profit-driven owner of COMPAS helps in interpreting the purposes behind such an approach. The company's response contains several main rhetorical strategies to appropriate control over the specific discourse about COMPAS, as well as the more general discourse about algorithmic risk assessment.

[71] Discourse can actively engage in ideological work, whereby it acts as a form of social action in power relations in mediating the link between text and society. van Dijk, *Critical Discourse Analysis*, *supra* note 41, at 467.

[72] Dieterich et al., *supra* note 8, at 1.

[73] *Id.*

[74] Christina Kakavá, *Discourse and Conflict*, *in* HANDBOOK OF DISCOURSE ANALYSIS 650, 653-54 (Deborah Schiffrin et al. eds., 2008) (quoting Deborah Schiffrin, *Everyday Argument: The Organization of Diversity in Talk*, *in* 3 HANDBOOK OF DISCOURSE ANALYSIS 35, 37 (Teun van Dijk ed., 1985)).

a. *Strategy One: The Choice of Classification Error*

Among Northpointe's specific discursive attacks, the one that appears at the heart of the conflict concerns the choice of classification errors in assessing bias. to better understand what this scientific dialogue is about, a visual may be useful. An industry-standard table exists for reporting on the utility of a categorical model that uses a binary classifier and a binary outcome — for purposes here, this entails a classifier of high risk (yes/no) and an outcome of committing a recidivist act (yes/no). The model is called a contingency table and is represented in Figure 1.

Figure 1: A 2 × 2 Contingency Table

		Outcome		
		Recidivist	Non-Recidivist	
Tool Result	Predicted to Recidivate	True Positives (TP)	False Positives (FP)	*Positive Predictive Value (PPV)*
	Not Predicted to Recidivate	False Negatives (FN)	True Negatives (TN)	*Negative Predictive Value (NPV)*
		False Negative Rate (FNR)	*False Positive Rate (FPR)*	

A contingency table is useful to calculate at least four different classification statistics. Two of these are the False Positive Rate ("FPR") and the False Negative Rate ("FNR"), which are derived vertically by column; they represent retrospective measures in which risk predictions from a tool are observed for the groups of known recidivists and non-recidivists, respectively. The FPR asks: of those who were not recidivists, what percentage of them were erroneously classified as high risk? The FNR asks: of the recidivists, what percentage of them were incorrectly classified as low risk? Alternative terminology includes *specificity*, which is the reciprocal of the FPR (i.e., 1 – FPR), and *sensitivity*, which is the reciprocal of the FNR (1 – FNR). The definitions of sensitivity and specificity will be useful in later discussion.

In contrast, the Positive Predictive Value ("PPV") and the Negative Predictive Value ("NPV") are calculated horizontally. These are predictive in nature in that they focus on the groups predicted by the tool to recidivate or not, and then calculate the percentage who actually recidivated or not, respectively. Hence, the PPV asks: of the persons testing positive (i.e., high risk), what percentage of them recidivated? NPV asks: of persons testing negative (i.e., low risk), what percentage of them did not recidivate?

The equations in Figure 2 apply to the foregoing classification statistics:

Figure 2: Classification Equations

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \qquad \mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}$$

$$\mathrm{Specificity} = 1 - \mathrm{FPR} \qquad \mathrm{Sensitivity} = 1 - \mathrm{FNR}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad \mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

Contingency table classification statistics require that the sample be divided into two groupings: one representing individuals predicted to be recidivists and the other non-recidivists. The discourses here generally use the COMPAS decile score of 5 as the cut point for this dichotomy, such that deciles of 1–4 are designated low risk and deciles 5–10 as high risk. Unless otherwise stated, the classification equation results use this cut point 5.

Importantly, the reason to clearly distinguish the FPR/FNRs, on the one hand, from the PPV/NPVs, on the other hand, for the Broward County dataset becomes evident. According to ProPublica, the FPR for blacks versus whites was 45% and 24%, respectively, while the FNR for blacks versus whites was 28% and 48%, respectively.[75] These differentials are large: for blacks the FPR is 21 percentage points higher than whites and its FNR is 20 percentage points lower than whites. In other words, the tool on these measures overestimated recidivism for blacks but overestimated non-recidivism for whites.

In contrast, according to Northpointe, the PPV for blacks versus whites was 63% as compared to 59%, respectively.[76] And the NPV for

[75] Angwin et al., *supra* note 5.

[76] Dieterich et al., *supra* note 8, at 20.

blacks was 65% compared to 71% for whites.[77] The rate differentials here are obviously much smaller, whereby for blacks the PPV is 4 percentage points higher than whites and the NPV is 6 percentage points lower. Thus, the tool was more accurate at predicting recidivism for blacks, but better at predicting non-recidivism for whites.

Overall, Northpointe's preference for PPVs and NPVs is pragmatic in serving its own interests.[78] The smaller PPV/NPV differentials serve to downplay racial contrasts. Moreover, the single digit disparities in PPV/NPVs seem to empower Northpointe to assert that these are "evidence of predictive parity" of the scale for blacks and whites.[79]

Northpointe uses discursive rationalization for this assertion. The company's overall rhetorical choice is to conceptualize the four classification statistics in the contingency table into two sets. The company labels the FPRs and FNRs as "Model Errors."[80] The PPVs and NPVs are accuracy statistics, and thus to render them compatible as error terms, their complements are used, as in 1 – PPV and 1 – NPV, respectively. Northpoint therefore labels these latter complementary terms as "Target Population Errors."[81] The distinction underlies Northpointe's discounting the FPR/FNR statistics which supported racial disparities in ProPublica's report (i.e., "Model Errors") in favor of the PPV/NPV metrics that indicate racial equity and promoted by Northpointe (i.e., "Target Population Errors," which, again, are the complements of PPV and NPV).

It is notable that in its argumentation on this matter, Northpointe draws on what the CDA literature refers to as the strategic employment of "implications" and "presuppositions."[82] These define semantic presentations of power that surreptitiously put forward concepts as if they were recognized facts, but that may actually not be true.[83] Here, Northpointe's discourse submits "Model Errors" and

[77] *Id.* at 20-21.

[78] *See* Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 101 (2017) (noting algorithmic tool developers have a natural incentive to make choices that improve predictive ability that may otherwise diverge from societal interests or contradict criminal justice policy); Andrew D. Selbst, *Disparate Impact in Big Data*, 52 GA. L. REV. 109, 136 (2017) ("Data brokers' incentives are to make their models just good enough so that their customers can profit more by using them than by not using them.").

[79] Dieterich et al., *supra* note 8, at 20-21.

[80] *Id.* at 7-8.

[81] *Id.* at 8.

[82] van Dijk, *Critical Discourse Analysis*, *supra* note 41, at 473.

[83] *Id.* at 473.

"Target Population Errors" as if they were terms of art in the algorithmic data science industry by repeatedly emphasizing them as capitalized terms and often in italics to catch the eye and bolster their seeming importance. Its pretention is that these are not actually industry recognized terms. Still, the strategy seems as though it is meant to bolster Northpointe's legitimacy. In CDA terms, this is a form of personal authorization:[84] Northpointe crafts original terminology to reorient the language of the conflict, yet without acknowledging its own engineering and without sufficient external confirmation. Indeed, a review of publicly available documents about COMPAS authored by Northpointe scientists fails to uncover any time prior to the conflict with ProPublica that these authors used the terms "model error" or "target population error."[85]

Nonetheless, Northpointe rationalizes the limits of the "Model Errors" as follows:

> *Model Errors* are of no practical use to a practitioner in a criminal justice agency who is assessing an offender's probability of recidivating. The practitioner does not know at the time of the assessment if the offender is a recidivist or not. *Model Errors* cannot be directly applied to an offender at the time of assessment.[86]

Northpointe argues that the PPV/NPVs, instead, have "clinical value" and "predictive value" because they are prospective in nature.[87]

Northpointe's discourse deploys further disparaging words to undermine ProPublica's scientific competence in the matter. For instance, Northpointe proclaims that ProPublica's reliance on the "Model Errors" were "mistakes."[88] Shortly thereafter, Northpointe

[84] van Leeuwen, *Discursive Construction*, *supra* note 48, at 106.

[85] *See, e.g.*, NORTHPOINTE, PRACTITIONER'S GUIDE TO COMPAS CORE (2015), http://epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-COMPASPractionerGuide.pdf; WILLIAM DIETERICH ET AL., PREDICTIVE VALIDITY OF THE COMPAS REENTRY RISK SCALES (2013), https://epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-MDOC_ReentryStudy082213.pdf (presenting results on an outcome study for the Michigan Department of Corrections); NORTHPOINTE, COMPAS SCALES AND RISK MODELS VALIDITY AND RELIABILITY (2010), https://epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-COMPASSummaryResults.pdf; Tim Brennan et al., *Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System*, 36 CRIM. JUST. & BEHAV. 21 (2009); TIM BRENNAN ET AL., RESEARCH SYNTHESIS RELIABILITY AND VALIDITY OF COMPAS (2007), www.northpointeinc.com/files/research_documents/reliability_validity.pdf.

[86] Dieterich et al., *supra* note 8, at 7 (citation omitted).

[87] *Id.* at 8.

[88] *Id.* at 6.

moralizes again, declaring that Propublica "misused" the FPR/FNRs as evidencing racial bias.[89] Even more forthrightly, Northpointe affirmatively, and in more judgmental terms regarding ProPublica, states: "They were wrong in doing that."[90]

In so arguing, Northpointe engages in the legitimization practice of personal authorization again, here in the implied form of "because I say so."[91] The company offers little external support for exalting the PPV/NPV metrics while censuring the FPR/FNRs. Indeed, as will be discussed further below, the position is contrary to industry standards, as well as Northpointe's own position in at least one other paper, which the company does not acknowledge.

b. Strategy Two: The Choice of Accuracy Equity Measure

After justifying its preferred statistics of the PPV/NPVs to refute the claim of racial bias, Northpointe then seeks to exhibit COMPAS's accuracy through a statistic known as the "area under the curve: ("AUC"). Northpointe characterizes the AUC as "one of the most widely used measures of diagnostic accuracy."[92] This rationalization to promote the AUC is in CDA language a type of "authority of conformity" based on its being a popular practice (as in "everybody's doing it").[93] The company acknowledges that the AUC is derived from a statistical plotting of true positive rates and false positive rates across a risk tool's scores.[94] In other words, the AUC is based on the FPR and the complement of FNR (i.e., the true positive rate = 1 – FNR) statistics.

Regarding the dataset underlying the debate, Northpointe states that the AUC results on the COMPAS general recidivism scale were .69 for blacks and for whites separately, thus declaring the tool to be "equally accurate for blacks and whites (equal discriminative ability) . . . [thus exhibiting] *accuracy equity*."[95]

The ability to claim — and to highlight by using italics — accuracy equity is important for Northpointe to legitimize its tool. However, in so doing, Northpointe's discourse engages in an internal contradiction. In denouncing ProPublica's use of FPR/FNRs to show differential

[89] *Id.* at 7.

[90] *Id.*

[91] *Discursive Construction*, *supra* note 48, at 106.

[92] Dieterich et al., *supra* note 8, at 7.

[93] *Discursive Construction*, *supra* note 48, at 109.

[94] Dieterich et al., *supra* note 8, at 23.

[95] *Id.* at 3.

utility for blacks, Northpointe declares that those statistics were wrong and incorrect classifications, and without clinical or predictive value. However, these same statistics are the sole values used to compute the AUC. As Northpointe itself points out when specifically discussing the AUC, FPR/FNRs are "useful for summarizing the accuracy of a risk scale"[96] as they "quantify *Model Error*" while their complements of specificity and sensitivity, respectively, "quantify *Model Accuracy*."[97] A related inconsistency is that Northpointe derogates ProPublica's promotion of statistics that were not predictive, yet it promotes the AUC despite the fact that the AUC is not forward-looking in nature either.[98]

The evident purpose of these contradictions is that Northpointe is attempting to rationalize two discordant positions. In sum, Northpointe wishes to (a) repudiate the values of FPR/FNRs presumably as they indicate racial disparity in that the rates are significantly different for blacks and whites, while at the same time (b) emphasizing that the AUC statistic — despite its reliance upon FPR/FNR-related statistics — shows comparable accuracy for blacks and whites. Yet these goals represent an internal contradiction and a lack of comprehensibility, plus are further complicated in the next strategic rendering.

c. Strategy Three: The Base Rate Problem

Northpointe's additional argument in elevating the so-called "Target Population Errors" over "Model Errors" regards base rate differences. Here, this means that the recidivism rates are significantly different. In the Broward County dataset, 52% of blacks recidivated, compared to 39% of whites. According to Northpointe, "*Model Errors* . . . ignore the base rate of recidivism."[99] The company further explains that "*Model Errors* . . . are calculated separately for recidivists and non-recidivists and . . . ignore the base rates for blacks and whites."[100] Hence, Northpointe contends that "ProPublica focused on classification

[96] *Id.* at 7.

[97] *Id.* at 33.

[98] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 PSYCHOL., PUB. POL'Y & L. 427, 431 (2016) ("As purely a retrospective discrimination index (i.e., distinguishing which reoffenders were previously determined to be high or low risk), the AUC does not deliver a forward-looking predictive estimate (i.e., forecasting which participants will actually go on to reoffend).").

[99] Dieterich et al., *supra* note 8, at 10.

[100] *Id.* at 10.

statistics that did not take into account the different base rates of recidivism for blacks and whites" and as a result such classification "statistics resulted in false assertions in their article."[101] In contrast, Northpointe highlights that its analysis is the true portrayal of COMPAS's abilities whereby the "Target Population Errors" (utilizing the complements of PPV and NPV) are the "correct classification statistics" by accounting for base rate differences between groups.[102]

However, Northpointe contradicts itself again. Recall that the AUC is derived from FPR/FNRs, specifically Sensitivity and 1 – Specificity.[103] When Northpointe discusses the AUC specifically, it inexplicably argues that "Sensitivity and Specificity can change as the base rate changes" and that "false positive rates increase[] with increasing base rate."[104] Evidently, this change of position was to allow Northpointe to account for the differences in sensitivity and specificity between blacks and whites as it at one point blames the disparities on the differences in base rates between the groups: "Differences in the base rates of blacks and whites for general recidivism (0.51 vs. 0.39) . . . in the [ProPublica] samples strongly effected the Sensitivity and Specificity tradeoffs observed in the [ProPublica] study."[105] Still, Northpointe's dueling strategies can further be understood by revealing certain omissions in Northpointe's response.

3. Omissions

Discourse analysts are also interested in what is absent or somehow missing from the rhetoric.[106] Herein are noted four examples of significant omissions in Northpointe's monologue that signify taking a disputable position without so conceding.[107]

[101] *Id.* at 1. In another part of the document, Northpointe basically repeats the point: "Obviously these model statistics (operating characteristics) do not depend on the base rate of recidivism." *Id.* at 33.

[102] *Id.* at 2.

[103] *Supra* Section III.A.2.a (Figure 2: Classification Equations).

[104] Dieterich et al., *supra* note 8, at 33 (citing Mariska Leeflang et al., *Variation of a Test's Sensitivity and Specificity with Disease Prevalence*, 185 CANADIAN MED. ASS'N J. E537, E539-40 (2013)).

[105] *Id.* at 7-8.

[106] van Leeuwen, *Moral Evaluation*, *supra* note 53, at 140.

[107] *See supra* text accompanying note 74.

a. *Conflict of Interest*

It is true that the authors of the Northpointe document do not hide their affiliation. Still, in the text itself the authors fail to clearly concede their conflict of interest in defending COMPAS. This might be understandable if the document purported to be a marketing brochure. But the discourse is couched in scientific and academic terms, and the audience may reasonably expect that statistics are wielded in an objective manner. Among the various critiques involving misrepresentations and omissions outlined herein, the unaddressed conflicts are telling. In CDA terms, it suggests a conscious hegemonic communications style by using deceptive practices in an attempt to control the dialogue about its risk tool.[108] In a sense, the silence on this conflict of interest signifies a reactionary retort to a threat to Northpointe's authoritative stance and the legitimacy of its prize tool.

One of Northpointe's positions in its response to ProPublica is to assert that COMPAS is based on "unbiased scoring rules."[109] Yet ProPublica reports that a Northpointe spokesman admitted that a risk assessment tool would likely have to include factors that were correlated with race as otherwise the tool's accuracy would decline.[110] Any tangible reduction in performance cannot much bolster profits for the tool's owner.[111]

b. *General Acceptance of ProPublica's Classification Statistics*

Despite Northpointe's disavowal of FPR/FNRs, such classification statistics are widely accepted in the data science and criminological communities for assessing the diagnostic accuracy of algorithmic risk tool abilities.[112] Indeed, many experts indicate that any equation

[108] *See* Wall et al., *supra* note 52, at 263.

[109] Dieterich et al., *supra* note 8, at 8.

[110] *See* Angwin et al., *supra* note 5 (emphasizing Brennan, the original creator of COMPAS, said "it is difficult to construct a score that doesn't include items that can be correlated with race — such as poverty, joblessness and social marginalization. 'If those are omitted from your risk assessment, accuracy goes down[.]'").

[111] *See* David Madras et al., *Learning Adversarially Fair and Transferable Representations* 1 (Oct. 22, 2018) (unpublished manuscript), arxiv.org/abs/1802.06309.

[112] *See, e.g.*, Paolo Eusebi, *Diagnostic Accuracy Measures*, 36 CEREBROVASCULAR DISEASES 267, 268 (2013) (accuracy of diagnostic medical tests); Christopher P. Marett & Douglas Mossman, *From Ballpark to Courtroom: How Baseball Explains Risk Assessment*, 47 PSYCHIATRIC ANNALS 443, 445 (2017) (illustrating the concept with statistics regarding baseball umpires' efficacy); Shaffi Ahamed Shaikh, *Measures Derived from a 2 x 2 Table for an Accuracy of a Diagnostic Test*, 2 J. BIOMETRICS & BIOSTATISTICS 1, 2 (2011) (accuracy of diagnostic medical tests); Karlijn J. van Stralen

commonly derived from the 2 × 2 contingency table (see Figure 1) qualifies in appraising a tool's classification ability.[113] A recent paper reciting the common definitions of algorithmic fairness counted the times each of them were cited in relevant literature and found that FPR/FNRs were cited almost twice as often as PPV,[114] confirming the general acceptance of these measures in the current algorithmic science world.

Northpointe's discursive approach to neglect this reality likely is a strategic one in its attempt to assert authority over the public understanding of its risk tool. The ploy could alternatively be characterized as a discursive misrepresentation to the extent that Northpointe expressly and repeatedly avers in its response to ProPublica that FPR/FNRs are inappropriate classification measures.[115]

Northpointe's position is more suspicious because of an assertion in a document it produced more recently. In its 2017 Practitioner's Guide to COMPAS, Northpointe reviews various validation studies using other datasets in supporting the utility of COMPAS.[116] In that document, the company clearly highlights that similar FPR and FNR rates (for the latter, using the complement of FNR) exhibited in a county sample in California were "providing evidence of *model error fairness*" which helped "demonstrate differential validity and fairness."[117] This positive use of FPR/FNRs contradicts Northpointe's contentions aimed at ProPublica.

Indeed, FPR/FNRs are known in the data justice literature on algorithmic fairness definitions as representing "equalized odds," meaning that equivalent FPRs and FNRs between groups signify that

et al., *Diagnostic Methods I: Sensitivity, Specificity, and Other Measures of Accuracy*, 75 KIDNEY INT'L 1257, 1259 (2009) ("The sensitivity, specificity, PPV, and NPV together result in four different measures, each indicating the accuracy of the test. All these measures have different pros and cons, and they may be difficult to interpret. Therefore, one sometimes prefers a combination of them." (internal citations omitted)).

[113] *See, e.g.*, Alexandria Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 155 (2017); Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* 5-6, 11 n.12 (Aug. 14, 2018) (unpublished manuscript), https://arxiv.org/pdf/1808.00023; Sahil Verma & Julia Rubin, *Fairness Definitions Explained* 3 (2018) (unpublished manuscript), http://www.ece.ubc.ca/~mjulia/publications/Fairness_Definitions_Explained_2018.pdf.

[114] Verma & Rubin, *supra* note 113, at 2 tbl.1.

[115] *See supra* Section IV.A.2.

[116] EQUIVANT, PRACTITIONER'S GUIDE TO COMPAS CORE (2017), http://equivant.volarisgroup.com/assets/img/content/Practitioners_Guide_COMPASCore_121917.pdf.

[117] *Id.* at 18.

the odds of such errors are equalized, and thus meet this definition of equity.[118] FPR/FNRs are also referred to under the algorithmic fairness idea named "error rate balance," whereby FPRs represent "false positive error rate balance," while FNRs are "false negative error rate balance."[119] Experts have characterized FPR/FNR differentials involving a protected group as producing disparate mistreatment.[120]

Thus, the implied portrayal that FPR/FNRs are inappropriate is not supported in the data science communities. From a legal perspective, too, it is unreasonable to discount FPR/FNRs. If either type of error disproportionately is visited on any particular group, disparate mistreatment may exist.[121] Hence, Northpointe's soliloquy on the irrelevance of any of the classification statistics is exaggerated and constitutes an attempt to obscure that its position here is disputable, if not firmly disputed, in the relevant scientific and legal communities.

At the same time, as will be discussed further below, FPR/FNRs and PPV/NPVs are merely a few of the algorithmic fairness measures that now exist in the data science literature.

c. Statistical Significance

Another omission in Northpointe's response should be noted. Northpointe purports to show predictive parity for blacks and whites through the use of PPV/NPVs, arguing that these statistics "refute" ProPublica's claim of racial bias.[122] Northpointe cites the PPVs for general recidivism as 63% for blacks and 59% for whites; the NPVs as 65% for blacks and 71% for whites.[123] Apparently, Northpointe implies that the differentials between the group rates on both statistics are practically similar, as it characterizes the differentials as being "slight[]."[124]

However, it is common practice in the statistics world to test the statistical significance of any differences in proportions between

[118] *See* Miconi, *supra* note 9, at 1.

[119] Verma & Rubin, *supra* note 113, at 2-4.

[120] *See, e.g.*, Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 CUMB. L. REV. 67, 95 (2018); Muhammad Bilal Zafar et al., *Fairness Beyond Disparate Treatment and Disparate Impact: Learning Classification Without Disparate Mistreatment*, INT'L WORLD WIDE WEB CONF. (2017), https://people.mpi-sws.org/~gummadi/papers/disparate_mistreatment_www2017.pdf.

[121] Geoff Pleiss et al., *On Fairness and Calibration* 2 (unpublished manuscript) (internal citation omitted), https://arxiv.org/pdf/1709.02012.pdf.

[122] *See* Dieterich et al., *supra* note 8, at 2, 20.

[123] *See id.* at 20-21.

[124] *See id.* at 11.

independent groups using a *z*-test.[125] Northpointe does not concede this practice in its response and thus does not provide readers with such information. One reason could be that when one performs this test, the results are not in COMPAS's favor. Perhaps appearing a bit prematurely, results from the quantitative component of this Article might still be useful here. Calculations of *z*-tests on the PPV and NPV differences between blacks and whites on the COMPAS general recidivism scale using the Broward County sample indicate they are both statistically significant (for the PPV, $p < .01$; for the NPV, $p < .0001$).[126] Here, then, Northpointe exaggerates the lack of meaningful difference between the groups on these measures as shown by these significantly small p values.

In sum, Northpointe's preferred metrics for algorithmic fairness illustrate disparate impact on blacks based on the NPVs. These results counter Northpointe's conclusion as, statistically speaking, there is not predictive parity for blacks and whites using Northpointe's preferred metrics for algorithmic fairness.

d. Additional Measures of Algorithmic Fairness

Northpointe designed much of the debate surrounding the choice of classification statistics and, during the course of which, attempted to socially construct ProPublica as intentionally skewing the results. The document recounts that the ProPublica "authors selectively reported and interpreted only the statistics that they thought supported their claim of racial bias against blacks."[127] Further, ProPublica "used the incorrect classification statistics to frame the COMPAS scales as biased against blacks."[128] Notice the use of the word "frame" to likewise connote deviant intention on ProPublica's part.[129] Indeed, in a more recent document, the Northpointe scientists relegate the ProPublica report to a "political context."[130]

[125] *See* RICCARDO RUSSO, STATISTICS FOR THE BEHAVIOURAL SCIENCES: AN INTRODUCTION 123 (2003) (indicating a *z*-test is appropriate to determine if proportions in two groupings are the same).

[126] The p value, which ranges from 0 to 1, indicates the significance of the test. A lower p value provides stronger evidence. Here, it would mean that as the p value approaches zero, the test is more strongly indicating a difference between two groups.

[127] Dieterich et al., *supra* note 8, at 24.

[128] *Id.* at 2.

[129] *See id.*

[130] Tim Brennan & William Dieterich, *Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)*, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 49, 64 (Jay P. Singh et al. eds., 2018).

Yet Northpointe's preference for classification statistics that better support the COMPAS tool is clearly self-serving considering its ownership interest.[131] Perhaps the strength of Northpointe's negative discourse here is to serve as a warning to ProPublica and others to keep their voices down, a strategy recognized by CDA researchers as a ploy to maintain a power imbalance.[132]

Notably, a significant gap in Northpointe's rhetoric here is ignoring the many other definitions of algorithmic fairness that are available in the relevant scientific literature. This omission may serve to skew the public's perception of the abilities of the COMPAS tool by not providing industry-standard, supplemental analyses. It also undermines the legitimacy of Northpointe's position by failing to consider available alternative perspectives. As a result, the next part of this Article intends to fill this gap by addressing various of the alternative algorithmic equity equations. These quantitative results may also be somewhat of a check on the interpretive critiques contained in the CDA above by being more transparent about contentious issues in the algorithmic risk assessment field.

B. Quantitative Results

The qualitative analysis contended that the major reason behind the seemingly inconsistent conclusions on racial bias is attributable to divergent options on classification statistics. The statistical explanation for this seeming anomaly was this: when base rates between groups vary, it is impossible to achieve equalized odds (FPR/FNRs) and predictive parity (PPV/NPVs) at the same time.[133] The reason is that PPV/NPVs are a function of the combination of equalized odds and base rates ("BR"), as reflected in the equations provided in Figure 3.

[131] *Cf.* van Dijk, *Critical Discourse Analysis*, *supra* note 41, at 466, 472-73 (discussing how speakers may manipulate discourse to serve their own interests).

[132] *See id.* at 471.

[133] *See* Miconi, *supra* note 9, at 2.

Figure 3: PPV and NPV Computations*

$$\text{PPV} = \frac{\text{Sensitivity} \times \text{BR}}{(\text{Sensitivity} \times \text{BR}) + ((1 - \text{Specificity}) \times (1 - \text{BR}))}$$

$$\text{NPV} = \frac{\text{Specificity} \times (1 - \text{BR})}{(\text{Specificity} \times (1 - \text{BR})) + ((1 - \text{Sensitivity}) \times \text{BR})}$$

* Where Sensitivity = 1 – FNR; Specificity = 1 – FPR; BR = base rate.

In the last few years, scholars in computer science, statistics, and criminology have developed various definitions of algorithmic fairness.[134] Some of these definitions overlap and may vary just by the name given by the particular field. Still, as with the example of the FPR/FNRs and the PPV/NPVs just given, several of the algorithmic fairness definitions are mutually exclusive or, as designated by statisticians, exemplify "impossibility theorems."[135] Due to the multiple, and at times incompatible, definitions available, a scholar has correctly observed that "any predictor can always be portrayed as biased or unfair, by choosing a specific measure of fairness."[136]

This Section advances additional algorithmic fairness definitions and provides relevant calculations. The point is to show how well COMPAS performs according to popular measures. We shall start with a rather straightforward standard.

1. Statistical Parity

Statistical parity exists when the percentages of offenders predicted to recidivate and those predicted not to recidivate are the same across groups.[137] More specifically, statistical parity is met when across groups these calculations achieve similar results as provided in Figure 4 (using the acronyms provided in the contingency table in Figure 1):

Figure 4: Statistical Parity Equations[138]

$$\text{Predicted yes} = \frac{\text{TP} + \text{FP}}{N} \qquad \text{Predicted no} = \frac{\text{TN} + \text{FN}}{N}$$

[134] Berk et al., *supra* note 36, at 12.

[135] *See id.* at 17.

[136] Miconi, *supra* note 9, at 4.

[137] *See* Berk et al., *supra* note 36, at 14.

[138] *Id.* at 13-14.

The literature also refers to this measure of equity as demographic parity,[139] equal acceptance rates, and group fairness.[140] Any significant difference is said to qualify as disparate impact[141] and adverse impact.[142]

Using the cut point of decile 5, COMPAS clearly does not achieve statistical parity as it predicts high risk at 58% for blacks compared to 33% of whites.[143] Still, this algorithmic fairness criterion has been criticized as it is impossible to achieve without some form of affirmative action-based intrusion when base rates vary,[144] as they do with blacks and whites in the Broward County sample.[145]

2. Differential Prediction

Differential prediction demonstrates group differences in predictive ability and its existence indicates predictive bias.[146] Researchers examining group bias in psychological testing in education have, with the endorsement of the American Psychological Association, standardized a methodology to empirically confirm its existence.[147] Group bias represents test bias; in turn test bias is present if systematic

[139] *See, e.g.*, James E. Johndrow & Kristian Lum, *An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction*, ANNALS APPLIED STATISTICS 3 (forthcoming), https://www.e-publications.org/ims/submission/AOAS/user/submissionFile/30728?confirm=1d6331c2.

[140] Chouldechova, *supra* note 33, at 155.

[141] *See* Johndrow & Lum, *supra* note 139, at 25.

[142] *See* Christopher M. Berry, *Differential Validity and Differential Prediction in Cognitive Ability Tests*, 2 ANN. REV. ORG. PSYCHOL. & ORG. BEHAV. 435, 437 tbl.1 (2015).

[143] In a separate analysis not reported in the text, the author found the disparity exists at the higher cut point of 8 as well, where 27% of blacks were predicted to reoffend compared to 11% of whites (we use the term "separate analyses" to indicate data runs that do not form part of the main findings and are additional yet derive from the same set of analyses by the author).

[144] *See* Miconi, *supra* note 9, at 4.

[145] *Cf.* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 686 (2017) ("While statistical parity seems like a desirable policy because it eliminates redundant or proxy encodings of sensitive attributes, it is an imperfect notion of fairness.").

[146] *See* Christopher D. Nye & Paul R. Sackett, *New Effect Sizes for Tests of Categorical Moderation and Differential Prediction*, 20 ORG. RES. METHODS 639, 640 (2016).

[147] Nathan R. Kuncel & Davide M. Klieger, *Predictive Bias in Work and Educational Settings*, *in* THE OXFORD HANDBOOK OF PERSONNEL ASSESSMENT AND SELECTION 462, 463 (Neal Schmitt ed., 2012) (confirming endorsements also from the National Council on Measurement in Education and the American Educational Research Association).

errors exist in how well a test measures members of different groups.[148] The gold standard for evaluating test bias involves a series of nested models of regression equations involving the test, the group(s) of interest, and an interaction term (test × group) as predictors of test outcomes.[149]

Notably, Northpointe is clearly aware that this method of testing for racial bias exists. In responding to ProPublica, Northpointe reflects that the "standard way to test for race effects is to fit a model with recidivism as the outcome variable and risk score, race, and race by risk score as predictors."[150] This description depicts the gold standard for evaluating test bias just mentioned. Yet the company inexplicably fails to conduct such an analysis, despite concluding that COMPAS shows no racial inequity. This is a glaring omission that calls for correction.

The nested models method detects group differences in the form of the relationship between the test and the outcome in terms of intercepts and slopes,[151] either of which reveals differential prediction.[152] The rule of thumb in the psychological assessment field is that a significant group difference in either the intercept or the slope represents that a single regression equation for the groups combined will predict inaccurately for one or both groups; in such a case, a separate equation for each group must be considered.[153] Unequal intercepts or slopes signify disparate impact,[154] without requiring evidence of any discriminatory intent.[155] Selected researchers in criminal justice have recently begun to apply this methodological

[148] *See* Adam W. Meade & Michael Fetzer, *Test Bias, Differential Prediction, and a Revised Approach for Determining the Suitability of a Predictor in a Selection Context*, 12 ORG. RES. METHODS 738, 738 (2009).

[149] *See* Jeanne A. Teresi & Richard N. Jones, *Bias in Psychological Assessment and Other Measures*, *in* 1 APA HANDBOOK OF TESTING AND ASSESSMENT IN PSYCHOLOGY 139, 144 (2013).

[150] Dieterich et al., *supra* note 8, at 19.

[151] *See* Jennifer L. Skeem & Christopher T. Lowenkamp, *Race, Risk, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 690-92 (2016).

[152] *See* Meade & Fetzer, *supra* note 148, at 740.

[153] *See* Cecil R. Reynolds & Lisa A. Suzuki, *Bias in Psychological Assessment: An Empirical Review and Recommendations*, *in* HANDBOOK OF PSYCHOLOGY 82, 101 (Irving B. Weiner ed., 2003).

[154] *See* Meade & Fetzer, *supra* note 148, at, 741 (2009).

[155] *See* Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109. 121-22 (2017).

practice of nested models to evaluate group bias in recidivism risk tools.[156]

The nested model structure here utilized variables labeled as Black (coded as black = 1, white = 0), the COMPAS decile score, and an interaction between them as Black × COMPAS decile score. A four-model structure is employed with the outcome variable being recidivism. The results are compiled in Table 1.

Table 1: Logistic Regressions Predicting the Odds of General Recidivism

Predictor	Model 1	Model 2	Model 3	Model 4
Black	1.710***	---	1.140*	1.187
Decile	---	1.327***	1.319***	1.327***
Black×Decile Interaction	---	---	---	0.991
Constant	0.642	0.239	0.227	0.222
-2LL	7209.06	6553.13	6548.77	6548.62
χ^2	89.35	745.28	749.64	749.79
n=5,278				

Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Coefficients represent odds ratios.

Model 1 signifies that the odds of recidivism for blacks are 71% higher than for whites with no controls. This finding is consistent with the higher base rate for blacks. Model 2 establishes the utility of COMPAS in that the odds of recidivism increase by 33% for every one decile increase in COMPAS score.

The higher intercept indicated in Model 3 for blacks is statistically significant. It signifies underprediction for blacks and demonstrates test bias. Though, the lack of a statistically significant interaction in Model 4 (i.e., Black × Decile) means that there is not bias in the slope. The Model 4 finding indicates that COMPAS decile scores carry basically the same strength of prediction as deciles increase for both blacks and whites. In sum, using the best practices model for test bias, COMPAS is positive for test bias based on race, though in a manner that advantages blacks. These results are relatively compatible with the PPV/NPV results trumpeted by Northpointe.

[156] *See, e.g.*, Jennifer Skeem et al., *Gender, Risk, Assessment, and Sanctioning: The Cost of Treating Women Like Men*, 40 LAW & HUM. BEHAV. 580, 585 (2016) (risk assessment of men and women).

3. Calibration

Calibration concerns absolute predictive accuracy in terms of how accurately a tool statistically estimates the outcome of interest.[157] A tool is well-calibrated in the first instance and then across groups "if the algorithm identifies a set of people as having a probability *z* of constituting positive instances, then approximately a *z* fraction of this set should indeed be positive instances. Moreover, this condition should also hold when applied separately in each group."[158]

For COMPAS, the tool is not well-calibrated for either group. At cut point 5, the tool predicted that 58% of blacks would reoffend, but only 52% did. The direction was the opposite for whites, whereby the tool predicted that 33% would reoffend, but 39% overall did.[159] These results indicate disparities in calibration ability based on race and that the algorithm is calibrated harsher for blacks than whites.

Data scientists indicate this result is a consequence of having similar PPV/NPVs whereby calibration tends to suffer in the face of group base rate differentials.[160] The algorithm is calibrated harsher on the group with the higher base rate, being blacks in this study.

One reason that the calibration result here shows bias against blacks, while the results of the test bias models in the differential prediction section indicated the bias there favored blacks, may be differences in the measurement level of the predictor: the former was based on a single cut point (high risk versus low risk) while the test bias method used scores across the ten deciles. In other words, the two equity definitions used a dichotomous and a continuous variable, respectively, of COMPAS scoring.

4. Mean Score Differences

An alternative algorithmic fairness condition references the concept of "balance for the positive class." It requires that the mean test score for those in the positive class — here, meaning recidivists — be the

[157] L. Maaike Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate Its Accuracy*, 44 CRIM. JUST. & BEHAV. 8, 11 (2017).

[158] Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, 2 (unpublished manuscript) (Nov. 17, 2017), https://arxiv.org/abs/1609.05807 (internal citations omitted).

[159] In a separate analysis not reported in the text, the author found that at cut point 8, the recidivism rates remain the same (52% or blacks and 39% for whites). But the predicted rates are lower than cut point 5. At cut point 8, 27% of blacks were predicted to reoffend compared to 11% of whites. Thus, unlike at cut point 5, the predictive rate was lower than the actual rate for blacks.

[160] *See* Miconi, *supra* note 9, at 3.

same across groups.[161] Correspondingly, "balance for the negative class" requires equal mean test score for those in the negative class — i.e., non-recidivists.[162] Table 2 presents mean COMPAS scores comparing recidivists and non-recidivists.

Table 2. Mean Decile Scores for the General Recidivism Scale

	Recidivists	Non-Recidivists
Blacks	6.24	4.22
Whites	4.72***	2.94***

Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Clearly, balances for the positive and negative classes fail. Black recidivists and non-recidivists receive significantly higher COMPAS scores on average than whites. Mean scores for recidivists and non-recidivist blacks are more than one decile higher than for whites within those same categories. Moreover, the mean score for non-recidivist blacks is closer to the mean score for recidivist whites, with a difference of only half a decile.

The higher recidivist mean score is consistent with the FNR being lower for blacks as the bar (mean risk score) is higher. Then the higher mean score for non-recidivist blacks is consistent with their higher FPR. Another way to convey this result is that as blacks have a higher expected base rate of reoffending, the tool will on average classify them with higher scores. As a result, blacks, both recidivists and non-recidivists will, as confirmed here, have higher mean scores.

5. Treatment Equality

Treatment equality considers the ratio of the errors, as in FN/FP or FP/FN, and thus is also known as the cost ratio of errors.[163] This fairness metric is not as popular as the others yet still has relevance. For blacks the FP/FN cost ratio is 1.4 and for whites it is 0.7. Hence, for blacks, false negatives are more costly than false positives, yet the opposite occurs for whites whereby false positives are more costly than false negatives. This means that blacks are being treated differently by the algorithm whereby a false negative is a bigger mistake for blacks. In other words, the algorithm is willing for blacks

[161] *See* Kleinberg et al., *supra* note 158, at 2 (emphasis omitted).

[162] *See id.* (emphasis omitted).

[163] *See* Berk et al., *supra* note 36, at 5, 15.

to have 140 false positives for every 100 false negatives, thus indicating over-classification for blacks.

If we look at the flipside, which is FN/FP, the cost ratio for whites is 1.4. The algorithm for whites is instead permitting 140 false negatives for every 100 false positives, here meaning under-classification of whites.

This definition of algorithmic fairness is decidedly in favor of whites over blacks: the algorithm is willing to assume a greater rate of false positives over false negatives for blacks, but the reverse is true for whites.[164] The results here are not too surprising because of another "impossibility theorem" involving the existence of unequal calibration. As previously reported, differential calibration existed whereby COMPAS predicted a higher percentage of blacks would reoffend than actually did while the opposite was true for whites. The trouble is that "it is effectively impossible to achieve calibration if the cost ratio of false negatives to false positives is not 1.0. Indeed, calibration only makes formal sense when the costs for both kinds of errors are the same."[165]

6. Limitations

Several limitations of the quantitative study should be mentioned. The single site limits generalization of results. This study relied upon archival data, and it is therefore possible for systematic errors to exist in data collection that are not observable on secondary data analysis. Recidivism outcomes were from official records and thus will not include undetected crimes. The dataset did not include interrater reliability scores that would confirm the dependability of COMPAS scoring across evaluators and over time. It would have been preferable to control for aspects of supervision as pretrial services and conditions may moderate reoffending rates, but such an option is also not available in this secondary data analysis.

CONCLUSION

The qualitative study reviewed two powerful players who discursively directed the attention of policymakers and members of

[164] In a separate analysis not reported in the text, the author found that the results are slightly different at cut point 8 whereby for whites there are more false negatives than false positives. The FN/FP ratios are 4.8 for blacks and 10.8 for whites. This still means at the higher cut point, false positives are far more costly for whites.

[165] Richard Berk, *Accuracy and Fairness for Juvenile Justice Risks Assessments*, 16 J. EMPIRICAL LEG. STUD. 175, 181 (2019).

the public regarding the potential for racial bias by a popular risk assessment tool. The investigative media group understood the political power and newsworthiness of a charge of racial bias in a criminal justice context. In turn, the risk tool's corporate owner attempts to mediate such dialogue by using its authoritative image in data science and through judgmental commentaries to undermine its rival's authority. Northpointe took aim at ProPublica's alleged statistical errors while also seeking to derail the publicity by reorienting the attack as politically-motivated.

Both groups bear some responsibility in selecting and emphasizing the algorithmic fairness definitions that most suited their perspectives. Supplemental algorithmic equity modeling performed herein demonstrates mixed results, in large part because of the impossibility theorems whereby certain algorithmic fairness models are incompatible with each other in real world settings. The gold standard for test bias is positive (bias in the intercept) though, in favor of blacks. Nevertheless, other equity measures (i.e., statistical parity, calibration, balance for the positive and negative classes, and treatment equality) strongly support disparate impact in the form of overprediction for blacks. By ignoring these additional models, Northpointe, as the owner of the COMPAS tool studied here, is particularly engaged in ideological manipulation to protect its assets. With a profit interest, it is an understandable position. Still, this exercise confirms the need for careful attention by lawyers, civil rights advocates, and data scientists, underscoring the benefit of third-party audits of nontransparent algorithms.

To be clear, this Article makes no conclusions as to whether algorithmic risk assessment tools should (or should not) be used in criminal justice decisions — though this query should not have a binary response in any event. Instead, it reiterates that algorithmic risk assessment tools, no matter how progressive, scientifically-informed, and algorithmically-sophisticated they may be, can still result in disparate impact. Hence, as civil rights groups and data scientists have recently warned, care must be taken with their use.