The Institutional Life of Algorithmic Risk Assessment

Alicia Solow-Niederman, YooJung Choi, and Guy Van den Broeck

Table of Contents

On August 28, 2018, California passed the California Money Bail Reform Act, also known as Senate Bill 10 (SB 10), and eliminated the state's system of money bail. Lauded by politicians as a "transformative" measure,[1] SB 10 aimed to remove wealth considerations from the administration of justice through a system of algorithmic pretrial risk assessment. Criminal justice reform groups, led by organizations such as the American Civil Liberties Union of Northern California, decried the text as enacted. They feared that despite the importance of criminal justice reform, this particular statute created "an overly broad presumption of preventative detention" that compromised due process and racial justice.[2] Within months, a separate political development intervened in this substantive debate: a coalition of bail bond industry groups engaged the state's direct democracy system and obtained the signatures required to qualify a referendum for the 2020 ballot, thereby staying the statute's implementation.[3]

Though SB 10 is thus moot for now, we cannot stay the bigger picture questions. Indeed, minding the gap between, on the one hand, politician's claims about its transformative effect on inequitable money bail practices and, on the other, civil society objections that the turn to algorithmic risk assessment replaces one flawed paradigm with another reveals systemic challenges around the use of these algorithmic tools.

Algorithmic pre-sentencing methods rely on the premise that an algorithm can provide a concrete measure of risk, inform a judge of salient facts about the defendant, and thereby make the system more objective and fairer for all.[4] But even if it makes sense to eliminate money bail, algorithmic pretrial assessments are not in fact an objective substitute for subjective human considerations. Building from a long-standing critique of actuarial assessments in criminal justice,[5] a rapidly-growing legal and technical literature recognizes that algorithms are not unbiased. Scholars, civil society members, and policymakers alike are contending with questions of bias and

---

[1] *See* Press Release, Governor Brown Signs Legislation to Revamp California's Bail System, Protect Public Safety, CA.GOV (Aug. 28, 2018), https://www.ca.gov/archive/gov39/2018/08/28/governor-brown-signs-legislation-to-revamp-californias-bail-system-protect-public-safety/index.html (quoting California State Chief Justice Tani Cantil-Sakauye: "This is a transformative day for our justice system. Our old system of money bail was outdated, unsafe, and unfair. It took a three-branch solution with Governor Brown, the Legislature led by Senator Hertzberg and Assemblymember Bonta, and the Judicial Council's Administrative Director Martin Hoshino working with judges in my Pretrial Detention Reform Work Group to bring about a fair and just solution for all Californians"). *See also id.* (quoting former Governor Jerry Brown: "Today, California reforms its bail system so that rich and poor alike are treated fairly").

[2] Press Release, ACLU of California Changes Position to Oppose Bail Reform Legislation (Aug. 20, 2018), https://www.aclusocal.org/en/press-releases/aclu-california-changes-position-oppose-bail-reform-legislation.

[3] *See* SB 10: Pretrial Release and Detention, CAL. CTS., https://www.courts.ca.gov/pretrial.htm (last visited Feb. 5, 2019); *see also* Jazmine Ulloa, *Bail Bond Industry Moves to Block Sweeping California Law, Submitting Signatures for a 2020 Ballot Referendum*, LA TIMES (Nov. 20, 2018, 4:05 PM), https://www.latimes.com/politics/la-pol-ca-bail-referendum-signatures-20181120-story.html.

[4] *See supra* note 1.

[5] As used here, "actuarial" refers broadly to empirically-informed assessments, and thus contrasts with judgments based only on clinical judgments made by professionals. *See* discussion *infra* text accompanying notes 22–30. *Cf.* NATHAN JAMES, CONG. RESEARCH SERV. R44087, RISK AND NEEDS ASSESSMENT IN THE FEDERAL PRISON SYSTEM 10 (2018) ("[I]t is argued that utilizing actuarial rather than clinical (i.e., professional judgment alone) risk assessment makes the process more objective and less susceptible to rater bias.").

accountability,[6] competing definitions of "fairness" within such algorithms,[7] and concerns about the ways in which automated tools may reinforce underlying societal inequities.[8] Research to date tends to focus on fairness, accountability, and transparency[9] within the tools, urging technologists and policymakers to recognize the normative implications of these technical interventions.[10]

While questions such as whether these instruments are fair or biased are normatively essential, this Essay contends that looking at these issues in isolation risks missing a critical broader point. Automated risk assessment systems do not operate in a vacuum; rather, they are deployed within complex webs of new and preexisting policy requirements as well as legal institutions, rules, and associated norms. To understand an algorithm's impact on human life and liberty, then, we must look at more than the instrument. Understanding its potential risks, affordances, and consequences requires us to grapple with the policymaking institutions and legal settings in which a given tool is embedded.[11]

This Essay's detailed analysis of SB 10 concretely illustrates how the terms of the algorithmic risk assessment regulation or statute interact with the tool's operation on the ground. Specifically, using SB 10 as a not-so-theoretical hypothetical reveals a tension between, on one hand, a top-down, *global* understanding of fairness, accuracy, and lack of bias and, on the other, a tool that is well-tailored to *local* considerations. There is a general conceit in the law that a principle like fairness is universal. Though what is needed to protect a right such as procedural due process might vary by

---

[6] *See, e.g.*, Vienna Thompkins, *What Are Risk Assessments — and How Do They Advance Criminal Justice Reform?*, BRENNAN CTR. (Aug. 23, 2018), https://www.brennancenter.org/blog/what-are-risk-assessments-and-how-do-they-advance-criminal-justice-reform; Anna Maria Barry-Jester et al., *The New Science of Sentencing*, MARSHALL PROJECT (Aug. 4, 2015, 7:15 AM), https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing.

[7] For an accessible overview of different technical definitions of fairness, see Arvind Narayanan, *Tutorial: 21 Fairness Definitions and Their Politics*, YOUTUBE (Mar. 1, 2018), https://www.youtube.com/watch?v=jIXIuYdnyyk.

[8] *See, e.g.*, Eric Holder, Former U.S. Attorney General, Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference (Aug. 1, 2014) (By basing sentencing decisions on static factors and immutable characteristics . . . [risk assessment tools] may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."), *available at* https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th.

[9] *See generally* Conference on Fairness, Accountability, and Transparency (FAT*), https://fatconference.org/index.html (last visited Feb. 4, 2019).

[10] By "normative," this Essay broadly refers to the effect of an intervention on the common good and/or the life and liberty of individuals. *Cf.* Laurence Solum, *Legal Theory Lexicon: Welfare, Well-Being, and Happiness*, LEGAL THEORY BLOG (May 31, 2009), https://lsolum.typepad.com/legaltheory/normative_legal_theory/ ("[A]ny or most of the reasonable views about normative theory agree that what is good or bad for individual humans is morally salient.").

[11] Legal scholars have recently decried a lack of concrete evidence about risk assessment tools' efficacy and called for a more practical, empirically-informed approach to evaluating risk assessment algorithms. *See* Megan T. Stevenson, Assessing Risk Assessment in Action, 3 MINN. L. REV. (forthcoming 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3016088 (manuscript at 3) [hereinafter Stevenson, *Assessing Risk Assessment*] (urging attention to the "people and design choices" behind risk assessment algorithms); (describing choices about risk prediction, the selection of a risk prediction algorithm, and division of group into classification levels as "choices that depend, at least partially, on the normative and legal landscape"); Brandon L. Garrett & John Monahan, *Judging Risk* (manuscript at 31-33) (Va. Pub. L. & Legal Theory Research Paper No. 2018-44, July 20, 2018), https://ssrn.com/abstract=3190403 (manuscript at 30) [hereinafter Garett & Monahan, *Judging Risk*] (arguing that "far more attention must be paid to the structure of decisionmaking that is supposed to be informed by risk assessment"). This Essay concurs that attention to decisionmaking structures and real-world outcomes, and not merely abstract risks, is imperative. It begins its human- and design-focused analysis one level up from the risk assessment tools themselves, and is the first account to evaluate how legislative and regulatory choices about a risk assessment regime create a particular institutional context within which tools operate—one that itself structures the affordances and limitations of the tools.

context,[12] the normative content of the right itself does not vary by jurisdiction. SB 10's text and legislative history support this globally-applicable perspective, calling for "validated risk assessment tools" that are "demonstrated by scientific research to be accurate and reliable."[13] Concepts like accuracy, reliability, and non-discrimination are fixed principles from such a legal and policy standpoint.

But this perspective clashes with the system of *algorithmic federalism* that we create each time a jurisdiction (typically a state) adopts risk assessment and then deploys the associated algorithmic risk assessment tool in its sub-jurisdictions (typically a county).[14] Reserving full exploration of algorithmic federalism and its structural features for future work, this Essay focuses on SB 10 to explore how a particular statute creates global and local tensions, along both technical and policy fault lines.

The design of a statute like SB 10 matters from both a technical and policy perspective. Consider, for instance, the discretion that SB 10 grants each county-level superior court ("SC"). According to the statutory text, each SC is to develop Pretrial Assessment Services ("PAS").[15] Each PAS is then to apply a risk assessment tool that it has selected from the list provided by the statewide Judicial Council ("Council").[16] Lacking, however, is trans-substantive guidance for decisions made at the county level. It is up to each county-level SC to make *local* decisions that give substantive meaning to any statutorily-stipulated first principles. As delineated in the Council's draft rules, the SC must determine "[w]hether any scientific research has raised questions that the particular instrument unfairly classifies offenders based on race, ethnicity, gender, or income level."[17] The problem is that there is no universal "scientific" agreement on what these concepts require,[18] nor even a rough consensus on minimal acceptable standards or procedures. Each county's court system must make the call. Accuracy, reliability, and non-discrimination are thus administered locally. Moreover, validating the tool requires careful consideration of local demographics, including how to contend with sensitive categories such as race and gender. Since each SC must ascertain "whether the particular instrument has been validated on a relevant population,"[19] determining what is or is not adequately reliable or accurate is again a local matter without top-

---

[12] *See* Mathews v. Eldridge, 424 U.S. 319, 424 (1976) ("Due process, unlike some legal rules, is not a technical conception with a fixed content unrelated to time, place and circumstances." (quoting Cafeteria Workers v. McElroy, 367 U. S. 886, 896 (1961))).

[13] See § 1320.7(g).

[14] Immense thanks to Kiel Brennan-Marquez for suggesting this term during an early presentation of the project. In future work (tentatively titled *Algorithmic Federalism*), Alicia Solow-Niederman intends to apply existing scholarship on federalism to explore local and global tensions, with an eye to the costs and benefits of allocating decision-making authority and discretion at different levels.

[15] See § 1320.26(a) ("The courts shall establish pretrial assessment services."); 1320.7(g) (defining "pretrial assessment services" as "entity, division, or program, at the option of the particular superior court").

[16] §1320.7(k) ("'Validated risk assessment tool' means a risk assessment instrument, selected and approved by the court, in consultation with Pretrial Assessment Services or another entity providing pretrial risk assessments, from the list of *approved pretrial risk assessment tools maintained by the Judicial Council*." (emphasis added)).

[17] Proposed rule 4.10(b)(5), http://www.courts.ca.gov/documents/SP18-23.pdf.

[18] *See* text accompanying *supra* notes 5–9 and sources cited therein.

[19] Proposed rule 4.10(b)(5), http://www.courts.ca.gov/documents/SP18-23.pdf.

down guidance.[20] Practically speaking, then, what law might like to be fixed becomes jurisdictionally variable.

The resulting challenges are not unique to SB 10; rather, anytime there is both a more centralized body that sets overarching guidelines about the tool and a risk assessment algorithm that must be tailored to reflect local jurisdictional conditions, there will be a *global-local tension*. For instance, the pursuit of a single, statewide understanding of a first principle like non-discrimination—as consistency and rule of law might demand—requires technical tradeoffs in the fairness and accuracy of the tool. Risk assessment tools must be validated with reference to the particular conditions of application.[21] Yet validation of a tool to make it more fair or accurate at a local level—as technological best practices demand—can produce different ground-level understandings of what "unfairly classif[ying]" or "reliable" requires. This local variability, however, is in tension with a top-down, global understanding of the normative principle.

This Essay uses SB 10 to explore these challenges and proceeds in four parts. Part I first surveys the adoption of actuarial criminal justice tools in the 20th century. It then canvasses recent state moves to implement algorithmic risk assessment tools as well as associated legal controversies and scholarly critiques. Next, Part II describes SB 10, elaborating on its provisions as a case study in the ways that a state might craft and deploy risk assessment instruments. Specifically, it summarizes particular SB 10 provisions and relevant legislative history, focusing on how the statute grants authority and discretion to institutional actors at both the state and local level. Part III applies this analysis with a series of hypothetical narratives drawn from real-world demographic data in California. These narratives illustrate how even the best-intentioned actions can lead to unanticipated and/or undesirable policy or technical results, given the way that a statute allocates authority to state, county-level, and court-level actors. Part IV extrapolates from these narratives to discuss the *local-global tension* in general. It closes by proposing that a policy choice to insert too many *layers of discretion* is likely to be problematic, no matter which tool is adopted, before offering several specific recommendations that could improve risk assessment statutes in general and SB 10 in particular.

---

[20] It is possible that Judicial Council rules around validation—development of which has been halted due to the 2020 ballot measure—would alter these dynamics. However, as discussed in the remainder of this Essay, they would need to do so by fixing global rules and foreclosing local tailoring, which introduces its own set of technical and policy concerns.

[21] *See, e.g.*, Garrett & Monahan, *Judging Risk*, *supra* note 11 (manuscript at 40) ("Instruments should be re-validated over time, at reasonable intervals, and with attention to local variation in populations, resources, and crime patterns." (internal citation omitted)); John Logan Koepke and David G. Robinson, *Danger Ahead: Risk Assessment and The Future of Bail Reform*, 93 WASH. L. REV. (forthcoming 2018) (manuscript at 25), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3041622 [hereinafter Koepke & Robinson, *Danger Ahead*] ("For tools to make well-calibrated predictions from the start, they need to be trained on data that matches the conditions about which they are making predictions."); PAMELA M. CASEY ET AL., OFFENDER RISK & NEEDS ASSESSMENT INSTRUMENTS: A PRIMER FOR COURTS 10, NAT'L CTR. STATE CTS. (2014) ("A local validation study will (a) inform any modifications that must be made to the content of the tool to optimize predictive validity in the local jurisdiction and ensure that it meets basic minimum scientific standards, and (b) inform the development of appropriate cutoff values for categorizing offenders into different risk.").

I. Risk Assessment Tools

A. Actuarial Justice: A Brief History

Contemporary risk assessment instruments share a common heritage with far older criminal justice interventions. The link between old and new is the idea that public officers can make predictions about an individual's behavior that should inform the treatment of that individual today. This Section's stylized overview contextualizes contemporary algorithms as the latest iteration of this historic phenomenon.

Beginning in the 1920s, parole boards increasingly invoked "crime prediction" in decisions about sentence length.[22] For much of the twentieth century, these choices about human liberty depended on subjective factors like "'the look in the prisoner's eye,' or [parole] board members' personal experiences, intuition, and biases."[23] And in making bail and sentencing determinations, "clinical predictions," or "the largely unstructured clinical judgment of skilled practitioners," were relied upon to assess the likelihood of recidivism. Outside of the parole or pretrial context, moreover, police officers and agencies have long made choices about where to allocate limited resources based on risk assessment,[24] an inherently predictive enterprise.

Two significant shifts occurred in the back half of the twentieth century. The first evolutionary arc was methodological. In the 1960s and 1970s, a growing sense that "clinical predictions" were unfairly subjective and hence susceptible to improper bias catalyzed "evidence-based" interventions.[25] These "second-generation" tools were "actuarial." Rather than rely on subjective expertise, they invoked statistics to "assign a quantitative risk score to an offender by assessing unalterable (e.g.[,] static) individual factors (i.e.[,] history of substance abuse and age at first offense) that have been statistically linked to the risk of recidivism in correctional populations and based on research involving large population samples."[26] In contrast to the paradigm shift from subjective to actuarial tools, subsequent developments have been more evolutionary. Over time, a "third generation" of statistical tools expanded beyond "static risk factors (such as criminal history, age, and gender)" to consider risks, needs, and "both static and dynamic risk factors such as educational status, and employment."[27] As Kelly Hannah-Moffat explains, tools that rely on more dynamic factors are distinct because they "focus on treatment or rehabilitation of the offender to

---

[22] *See* Thomas Mathiesen, *Selective Incapacitation Revisited*, 22 L. & HUMAN BEHAVIOR 455, 458–59 (1998), www.jstor. org/stable/1394595; Kevin R. Reitz, *"Risk Discretion" at Sentencing*, 30 FED. SENT'G REP. 68, 70 (2017) ("[P]rison sentence lengths in most U.S. jurisdictions are already based on predictions or guesses about offenders' future behavior, and this has been true—in multiple settings—for at least a century.").

[23] Reitz, *supra* note 22, at 69 (citing David J. Rothman, Conscience and Convenience: The Asylum and its Alternatives in Progressive America (1980); Marvin E. Frankel, Criminal Sentences: Law Without Order (1973)). *See also* Bernard Harcourt, Against Prediction 7–18 (2007) (discussing parole boards' use of risk assessment instruments since the 1920s).

[24] This Essay adopts a common definition of risk assessment as "the process of using risk factors [factors that precede and statistically correlate with recidivism] to estimate the likelihood (i.e., probability) of an outcome occurring in a population." *See* Garrett & Monahan, *Judging Risk*, *supra* note 11 (manuscript at 7 (citing Carnegie Commission on Science, Technology, and Government, *Risk and the Environment: Improving Regulatory Decision Making* (1993))).

[25] For discussion of studies reporting bias in human judgment, see Lauren Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, 20 CRIM. JUST. & BEHAVIOR 1, 17–18 (2018).

[26] Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, 30 JUSTICE Q. 270, 274 (2013). For a discussion of static tools' limitations, see *id.*

[27] Garrett & Monahan, *supra* note 11 (manuscript at 10). *See also* Hannah-Moffat, *supra* note 26, at 274–77.

prevent reoffending, rather than simply predict recidivism. . . . This approach to risk differs importantly from the correctional use of static risk for preventive or selective incapacitation, diversion, or deterrence of recidivism through the administration of harsh penalties."[28] The "fourth generation" of tools continued to use various combinations of static and dynamic inputs while according more weight to the individualized needs of the defendant.[29] The "fifth generation" entails the application of machine learning techniques, discussed below, to provide more up-to-date predictions that take far more factors into account.[30]

A second evolutionary arc occurred in parallel, driven in large part by political forces and associated policy shifts.[31] The early 1960s witnessed mounting concern with overcrowded jails and less wealthy defendants' detention, even when they did not pose any public safety risk if they were released.[32] These reform efforts culminated with Congress' enactment of the 1966 Bail Reform Act, which aimed to "minimize reliance on money bail[,] . . . established that a defendant's financial status should not be a reason for denying their pretrial release, made clear that the risk of nonappearance at trial should be the only criterion considered when bail is assessed, and . . . generally forbid judges from treating a defendant's dangerousness or risk to public safety as a reason for detention."[33]

But this initiative did not endure. In the 1970s and 1980s, the federal government undertook a fundamental reworking of the underlying reason for imposing bail.[34] Catalyzed by mounting public unrest in the civil rights era and the Nixon Administration's "tough on crime" stance, a second set of reforms effectively reversed the earlier policy. Rather than set terms that ensured a defendant's presence in court, pretrial detention emphasized the risk that the defendant would commit future crimes and threaten public safety.[35] In practice, as it emerged, this shift reframed the salient questions and asked judges to assess a defendant's perceived "dangerousness" to determine the risk they posed.[36] This evolution was formally codified in the Reagan Administration with Congress's enactment of the 1984 Federal Bail Reform Act, which required federal judges to "consider danger to the community in all cases in setting conditions of release."[37] States largely

---

[28] Hannah-Moffat, *supra* note 26, at 276. For a more detailed account of risk/needs assessment in more modern tools, *see id.* at 274–76.

[29] Garrett & Monahan, *supra* note 11 (manuscript at 10).

[30] Id.

[31] This overview is indebted to John Logan Koepke & David G. Robinson's summary of this history. *See Danger Ahead*, *supra* note 21 (manuscript at 8–15 and sources cited therein).

[32] *See id.* at 9.

[33] *Id.* at 9 (citing United States v. Leathers, 412 F.2d 169, 171 (D.C. Cir. 1969)).

[34] For a more detailed survey of changing bail practices in the U.S., see Koepke & Robinson, *Danger Ahead*, *supra* note 21 (manuscript at 6–20).

[35] Garrett & Monahan, *supra* note 21 (manuscript at 10). *See also* Koepke & Robinson, *supra* note 31 (manuscript at 10–13).

[36] Though beyond the scope of this paper, these developments occurred in tandem with a "selective incapacitation" movement that focused on detaining the most "dangerous" defendants. This effort dovetails with many of the objectives of recidivism-based risk assessment. For an analysis of contemporary lessons for algorithm risk assessment drawn from the history of selective incapacitation, see Danielle Kehl et al., *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing,* RESPONSIVE COMMUNITIES INITIATIVE, (manuscript at 3–16) (2017), http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041.

[37] *United States v. Himler*, 797 F.2d 156, 159 (3d. Cir. 1986). *See also* Koepke & Robinson, *supra* note 21 (manuscript at 14).

followed the federal government's lead, marking a national shift in the discourse around risk assessment.[38]

The move towards assessing "dangerousness" goes hand-and-hand with the ongoing evolution of predictive risk assessment instruments. To date, risk assessment tools rely on more simple machine learning (ML) models, such as logistic regression, and do not yet embrace more complex ML models.[39] ML operates by parsing large datasets to identify patterns in the data, which allows the development of a predictive function that can be applied to previously unseen data sets.[40] The use of an algorithm might be more standardized than the older clinical assessment based on, say, the look in the person's eye.[41] Yet it does not make the predictive enterprise objective; rather, it turns on a spate of human choices about what data to use, what statistical model to adopt, how to "tune" the model, and how to apply the findings.[42] As more complex ML methods are integrated into risk assessment instruments, it will become even more essential to resist "automation bias"[43] and ensure adequate oversight of the tool's fairness and accuracy. As this Essay underscores, the way that this predictive enterprise operates turns not only on individual choices, but also on policy decisions about how to constrain or channel discretion and decisional authority.

B. Algorithmic Risk Assessment: A Recent History

The practical stakes are high because state and local jurisdictions in the United States are increasingly turning to algorithmic risk assessment. Though existing implementations span the pretrial and sentencing context, this Essay focuses on pretrial risk assessment procedures like SB 10.[44] In the last seven years alone, half of U.S. states have either implemented or are seriously considering the use of some form of risk assessment tools in pretrial settings. And many of these developments have been quite recent. According to the National Council of State Legislatures (NCSL),[45] in 2017 alone, "nine states enacted laws allowing or requiring courts to use risk

---

[38] *See* Koepke & Robinson, *Danger Ahead*, *supra* note 31 (manuscript at 13–14).

[39] *Accord id.* (manuscript at 43) (suggesting that risk assessment is likely to progress "from logistic regression-based techniques toward more complex machine learning techniques"). Legal scholarship at times seems to distinguish between "real" machine learning and actuarial science based on logistic regression and other statistical methods. Rather than advance such a dichotomy, this Essay positions contemporary risk assessment algorithms based on logistic regression as a simple ML model.

[40] For discussion of machine learning, see generally VISHAL MAINI & SAMER SABRI, MACHINE LEARNING FOR HUMANS (2017), https://www.dropbox.com/s/e38nil1dnl7481q/machine_learning.pdf?dl=0. In supervised machine learning, the dominant contemporary method, the data scientist will "tune" different parameters to improve a selected statistical model's ability to deliver results that are closer to a predefined goal, or "objective function."

[41] *See* discussion *supra* text accompanying notes 22–24 and sources cited therein.

[42] For an overview of the different steps necessary to arrive at a working ML model, see David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 UC DAVIS L. REV. 653 (2017).

[43] *Cf.* Danielle Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1271–72 (2008) ("The impulse to follow a computer's recommendation flows from human 'automation bias'—the 'use of automation as a heuristic replacement for vigilant information seeking and processing.'" (quoting Linda J. Skitka et al., *Automation Bias and Errors: Are Crews Better Than Individuals?*, 10 INT'L J. AVIATION PSYCHOLOGY 85, 86 (2000))).

[44] This narrower focus permits more detailed analysis of a particular set of interventions, without inadvertently conflating pretrial risk assessment and other forms of algorithmic criminal justice decisions, such as sentencing. However, this Essay's broader conclusions about the importance of looking at the policy design, as well as its analysis of systems in which discretion is spread across "global" and "local" layers, are more generally applicable.

[45] This research is funded in part by the Laura and John Arnold Foundation, which also produces a widely-used Public Safety Assessment, or PSA, tool, that has been adopted or is being implemented in over 40 jurisdictions. *See* Pretrial Justice, LJAF, http://www.arnoldfoundation.org/initiative/criminal-justice/pretrial-justice/ (last visited Jan. 11, 2019). Though beyond the scope

assessments to assist in establishing bail and release conditions [and] [a]nother five passed bills directing studies or development of risk assessment tools."[46] These enactments are, moreover, the latest in a longer-running series of state statutes that use these tools: "Since 2012, 20 laws in 14 states created or regulated the use of risk assessments during the pretrial process. In 2014 alone, 11 laws were passed to regulate how risk assessment tools are used to help determine whether, and under what conditions, a defendant should be released."[47]

Significantly, these enactments represent more than updates to technical instruments in order to keep up with the Joneses: there is also an underlying policy narrative. Like earlier historic shifts, they reflect a belief that more sophisticated, algorithmic risk analysis tools can better account for individual defendant characteristics, rather than making a blanket choice to detain or release an individual based on their alleged charges.[48] And continuing debates that date to at least the 1960s, much of the discussion involves broader questions about the role of bail or non-monetary restrictions, as well as how these choices interact with fundamental rights and civil liberties. For instance, prior to California's elimination of money bail with SB 10, a 2017 New Jersey state statute changed New Jersey's money bail system to provide judges with algorithmic risk assessment scores, aiming to "build a better, fairer and safer system of criminal justice."[49]

This apparent state enthusiasm for algorithmic solutions, however, has met with mounting public and scholarly debate about the ethical and legal propriety of these tools. For instance, over a hundred civil society organizations recently signed and adopted a statement of civil rights concerns.[50] As this document details, issues with the use of such tools include, among others, the risk of data inputs that reproduce and reinforce racial inequities in the criminal justice system overall; a failure to provide adequate procedural safeguards, including individualized, adversarial hearings for all defendants; and a lack of transparency or access to the data or algorithms used by proprietary instruments.[51]

---

of this Essay, the lack of independent research or oversight of these tools by bodies that are not also invested in creating them is disconcerting. Cf. *The Use of Pretrial "Risk Assessment" Instruments:" A Shared Statement of Civil Rights Concerns*, http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf (manuscript at 7) [hereinafter *A Shared Statement of Civil Rights Concerns*] (last visited Jan. 11, 2019). ("If in use, a pretrial risk assessment instrument must be transparent, independently validated, and open to challenge by an accused person's counsel. The design and structure of such tools must be transparent and accessible to the public."); Garrett & Monahan, *supra* note 11 (manuscript at 40) ("The review of risk assessment instruments has not always been public, or by researchers independent of those who developed the instrument. If instruments are only validated in-house, or by the originators of the instrument, then the validations cannot be verified independently.").

[46] NCSL, *Trends in Pretrial Release: State Legislation Update Civil and Criminal Justice* 1 (Apr. 2018), http://www.ncsl.org/portals/1/ImageLibrary/WebImages/Criminal%20Justice/pretrialEnactments_2017_v03.pdf.

[47] Amber Widgery, *Trends in Pretrial Release: State Legislation*, NCSL 1 (2015) http://www.ncsl.org/portals/1/ImageLibrary/WebImages/Criminal%20Justice/NCSL%20pretrialTrends_v05.pdf.

[48] *See id.*

[49] *See* Stuart Rabner, New Jersey State Chief Justice, Opinion, *Chief Justice: Bail reform puts N.J. at the Forefront of Fairness*, NJ.com (Jan. 9, 2017, 9:33 AM), https://www.njcourts.gov/courts/assets/criminal/starledgercolumn.pdf.

[50] *A Shared Statement of Civil Rights Concerns*, *supra* note 45. *See also* Press Release, More than 100 Civil Rights, Digital Justice, and Community-Based Organizations Raise Concerns About Pretrial Risk Assessment (July 30, 2018), https://civilrights.org/more-than-100-civil-rights-digital-justice-and-community-based-organizations-raise-concerns-about-pretrial-risk-assessment/.

[51] *See id.* This Essay reserves treatment of due process concerns and questions about whether these algorithms amount to unlawful preventative detention for separate work.

These long-simmering issues, moreover, began to boil over into more general consciousness with a 2016 ProPublica investigation alleging that a proprietary sentencing algorithm, COMPAS, was systematically unfair in its treatment of black defendants.[52] Specifically, there was a problem with "error rate balance:" COMPAS labelled more black defendants who did *not* reoffend as "high risk" (false positives) and more white defendants who *did* reoffend as low risk (false negatives). This research was immediately met with rebuttals[53] contending that the tool is in fact unbiased because it exhibits "predictive parity:" at a given risk level, it predicted a roughly equal proportion of white and black reoffenders. Against this technical fairness debate, which turns on how one defines the concept, still others lodged critiques that no opaque, proprietary algorithm could be considered fair to a criminal defendant challenging it.[54] And others, including the company that produced the tool, argued that the problem in fact lies in the data itself, and not the algorithm. From this point of view, the bottom line is unequal base rates of recidivism among racial groups in the observed population that makes up the ML dataset. Given such baseline differences, any tool that applies proportionate outcomes across the board at a particular risk level will have proportionately different effects on different racial groups.[55] The debate about technical fairness thus connects to bedrock civil society concerns about racial inequity in the criminal justice system as a whole, even as the proponents of these tools assert that they are more objective, fairer ways to make criminal justice decisions.

Recent legal scholarship echoes these points. A rapidly-growing literature cautions against algorithmic risk assessment as an automatic key to a fairer criminal justice system. In addition to critiques of the biased racial impact of risk assessment[56] and evaluations of what "fairness" means in practice,[57] scholars have begun to assess critically the different ways in which algorithmic systems may be unfair. For example, a recent article by Lauren Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini proposes three "layers of bias" that can arise in risk assessment models: first, "challenges to fairness within the risk-assessment models themselves;" second, "biases embedded in data;" and, finally, whether it is fundamentally "fair to make criminal justice

---

[52] *See* Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. For discussion of "the COMPAS debate," see Eckhouse et al., *supra* note 25, at 6–7.

[53] *See* Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It's Actually Not that Clear.*, WASH. POST (Oct. 17, 2016) https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/. The company also authored a rebuttal. *See* William Dieterich et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Psredictive Parity*, NORTHPOINTE (July 8, 2016), https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html; *But see* Julia Angwin & Jeff Larson, *ProPublica Responds to Company's Critique of Machine Bias Story*, PROPUBLICA (July 29, 2016, 11:56 AM) https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story.

[54] *See, e.g.*, Matthias Spielkamp, *Inspecting Algorithms for Bias*, MIT TECH. REV. (June 12, 2017), https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/.

[55] For a detailed explanation of how algorithms can perpetuate society's historic racial inequities, see Sandra Mayson, *Bias In, Bias Out*, 128 YALE L.J. (forthcoming 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3257004.

[56] *See, e.g.*, Bernard E. Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment,* 27 FED. SENT'G REP. 237, 237 (2015) ("[R]isk today has collapsed into prior criminal history, and prior criminal history has become a proxy for race. The combination of these two trends means that using risk-assessment tools is going to significantly exacerbate the unacceptable racial disparities in our criminal justice system.").

[57] Again, there are myriad technical and ethical definitions that may be at odds with one another and with other values such as accuracy. For a video tutorial canvassing these issues, see Aaron Roth, *Tradeoffs Between Fairness and Accuracy in Machine Learning*, YOUTUBE (Jan. 30, 2017), https://www.youtube.com/watch?v=tBpd4Ix4BYM. *See also* text accompanying notes 18–20 and sources cited therein.

decisions about individuals based on groups" in the manner that algorithmic risk assessment demands.[58] This third prong questions whether it is constitutionally valid to make criminal justice choices in this way, raising both equal protection and due process concerns.[59] Adding to the dialog around constitutional principles and new technical interventions, Brandon Garrett and Jonathan Monahan have also raised constitutional questions involving judges' actual use of risk assessment algorithms, noting unsettled due process questions when a judge's determination is "informed by quantitative risk assessment methods."[60] As Garrett and Monahan acknowledge, and as Megan Stevenson contends in another recent piece,[61] despite a bevy of theoretical concern about the use of these algorithmic tools, there is an extremely limited literature on their adoption in practice.

There is not only a lack of empirical evidence about the use of algorithmic tools on the ground, but also even less sustained attention to the design of statutes and regulations and the associated protocols, norms, and institutions within which risk assessment instruments are developed and deployed. This Essay thus provides the first in-depth evaluation of ways in which choices about how to craft a policy and how to allocate discretion and authority within new and preexisting institutions inform both the theory and practice of algorithmic risk assessment.

II. Statutory Structure

SB 10 demonstrates how a statute's framework carries technical and legal implications for the operation of risk assessment instruments in the state. This Part surveys the statutory text as enacted, providing a foundation for Part III's analysis of the kinds of substantive outcomes that are possible within a given framework.[62] It is tailored to aspects of the statute that grant authority and discretion to different categories of institutional actors regarding the design and use of the pretrial risk assessment tool.[63]

On its face, SB 10 grants local courts considerable authority over the creation and implementation of the risk assessment instrument. The statute provides that local "Pretrial Assessment Services" are

---

[58] *See* Eckhouse et al., *supra* note 25, at 1.

[59] This branch of the literature relies on earlier work by scholars such as Sonja Starr, who has argued that actuarial risk assessments violate the Equal Protection Clause of the U.S. Constitution. *See* Sonja B. Starr, *Evidence-Based Sentencing and The Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014). *But cf.* sources cited *infra* note 103 (suggesting that contemporary Equal Protection jurisprudence is not an apt fit for algorithmic criminal justice).

[60] Garrett & Monahan, *supra* note 24, at 1–2.

[61] *See* Stevenson, *Assessing Risk Assessment*, *supra* note 11 (manuscript at 3).

[62] Though SB 10's implementation is stayed pending the November 2020 referendum, *see supra* text accompanying notes 1–4, this case study nonetheless acts as a rich, not-so-hypothetical example of the ways that structural and institutional choices, local demographics, and technical affordances and limitations interact, in context. An additional benefit, moreover, is that there is still time to alter the statutory and regulatory structure in response to the analytic and normative concerns explored here.

[63] This article integrates proposed rules that the Judicial Council has already published. In addition, this Essay highlights techno-policy ambiguities, such as precisely what "validation" demands. Though it is possible that additional rules might have clarified these matters, the ambiguities are unresolved for the foreseeable future because further rule development is stayed pending the 2020 referendum. *See supra* note 2. And the bottom line does not turn on this specific text. As explored *infra* Part IV, the very textual ambiguity and need for top-down rules to clarify the institutional application of the algorithm itself contributes to the "black box" policy regime created by layers of legislation, judicial rule-making, global and local technical and policy discretion, local entities' implementation of the tool, and local courts' use of the tool in particular cases.

responsible for pretrial risk level assessment of individuals who have been charged with a crime.[64] Each "particular superior court," which operates at the county level,[65] is to determine whether Pretrial Assessment Services ("PAS") consists of "employees of the court, or employees of a public entity contracting with the court for those services."[66] SCs may opt into a regional consortium or multi-county PAS with an "adjoining county," with the limitation that "persons acting on behalf of the entity, division, or program shall be officers of the court."[67] Moreover, there is local control over who makes up the "court," which the statute defines to include "'subordinate judicial officers,' if authorized by the particular superior court" in accordance with the California Constitution and Rules of Court.[68] SCs thus call the shots regarding creation of the PAS as a local institution.[69] This trial court control, moreover, continues after the creation of the PAS. Once the SC sets up a PAS, the PAS is to report this information to the trial court, along with recommended post-trial conditions of release.[70] These recommendations are non-binding, and discretion rests with the adjudicating judge.[71]

But this local discretion is not open-ended. Creating and using a risk assessment tool entails political and technical choices. And these decision points complicate the local control narrative because global statutory provisions and statewide Rules of Court interact with and constrain the available set of local choices. Specifically, SB 10 empowers the California Judicial Council ("Council"), which acts as the "policymaking body" for the California court system, to "adopt

---

[64] §§ 1320.26(a) ("The courts shall establish pretrial assessment services . . . [that] may be performed by court employees or the court may contract for those services with a qualified local public agency with relevant experience."); 1320.7(f) ("Pretrial risk assessment" means an assessment conducted by Pretrial Assessment Services with the use of a validated risk assessment tool."); 1320.7(g) ("Pretrial Assessment Services" means an entity, division, or program that is assigned the responsibility, pursuant to Section 1320.26, to assess the risk level of persons charged with the commission of a crime.").

[65] There are 58 trial courts in California, one per county, known as superior courts. These courts have general jurisdiction over all state civil and criminal matters, unless otherwise provided by statute or federal law. *See Superior Courts*, CAL. CTS., http://www.courts.ca.gov/superiorcourts.htm (last visited Jan. 12, 2019). The number of judges per superior court varies by the size of the county.

[66] § 1320.7.

[67] *Id.*

[68] § 1320.7(a) ("'The court' as used in this chapter includes 'subordinate judicial officers,' if authorized by the particular superior court, as authorized in Section 22 of Article VI of the California Constitution and specified in Rule 10.703 of the California Rules of Court.").

[69] Earlier versions of the statute called for a statewide oversight body to play a greater role in creation of the local PAS. *See* Assembly Committee on Public Safety (July 11, 2017) (manuscript at 35–36), *available at* http://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=201720180SB10# (describing an "unnamed agency" that would, among other responsibilities, be "authorized to oversee pretrial services agencies, to select a statewide pretrial assessment tool, to develop guidelines, and to provide training and assistance on pretrial release to judges, prosecutors, defense counsel, pretrial services agencies, jail staff, and law enforcement."). References to the "unnamed agency" and such top-down oversight were eliminated from the statute after the Chief Probation Officers of California expressed concern that SB 10 would "inhibit[] local control and flexibility relative to allowing each jurisdiction to determine who will handle the various parts of the pretrial program . . . at the local level." *Id.* 15–16. The Judicial Council also expressed concern at an earlier stage that SB 10 "would infringe on judicial discretion and independence." *Id.* at 13–14. Subsequent versions of the statute replaced the proposed "unnamed agency" with the final structure that combines the Council's oversight via rulemaking with increased discretion at the local level.

[70] § 1320.7(g).

[71] *See, e.g.*, §1320.20(f) ("Solely for the purpose of determining whether the person should be detained or to establish the least restrictive nonmonetary conditions of pretrial release to impose, the court may take into consideration any relevant information, as set forth in a California Rule of Court, including, but not limited to . . . [t]he recommendation of Pretrial Assessment Services obtained using a validated risk assessment instrument.").

California Rules of Court and forms" as needed to implement SB 10.[72] It is also to, among other responsibilities:

- "Compile and maintain a list of validated pretrial risk assessment tools;"[73]
- Identify, define, and collect "minimum required data to be reported by each court;"[74]
- "[T]rain judges on the use of pretrial risk assessment information when making pretrial release and detention decisions, and on the imposition of pretrial release conditions;"[75] and
- Consult with the Chief Probation Officers of California and "assist courts in developing contracts with local public entities regarding the provision of pretrial assessment services."[76]

The Council thus manages the development of validated risk assessment tools, top-down, and the decentralized PAS developed by each county-level SC applies the instruments in the specified manner. This allocation of authority might initially appear to emulate other contexts in an unproblematic way. The Council has long promulgated rules for the state's judicial system. Since its creation in 1926, it has operated under a state constitutional mandate "to improve the administration of justice."[77] In addition to general consideration for the public interest and judicial policymaking, the Council fulfils this mandate by setting official rules of court for the state.[78] The Council presently carries out its mission through a number of internal committees, advisory committees, and task forces, which generally include some combination of voting members and advisory members.[79] Moreover, in this particular context, the Council's centralized oversight of which tools are "validated" is arguably a good thing, lest jurisdictions employ a tool that is biased, discriminatory, opaque, or otherwise problematic.

Even if oversight is necessary, however, formally following the same rule-making playbook may produce different outcomes in different jurisdictions. Consider, for instance, the high/medium/low risk determination. This policy choice is a focal point at which SB 10's global rule-setting requirements interact with local discretion. According to the statutory text, the Council is to appoint a "panel of experts and judicial officers . . . [that] shall designate "low," "medium," and "high" risk levels based upon the scores or levels provided by the instrument for use by Pretrial

---

[72] § 1320.24.

[73] § 1320.24(e)(1).

[74] § 1320.24(b)(1); § 1320.24(e)(2).

[75] § 1320.24(e)(3).

[76] § 1320.24(e)(4).

[77] CAL. CONSTIT. art. VI, § 6. *See also* LARRY L. SYPES, COMMITTED TO JUSTICE: THE RISE OF JUDICIAL ADMINISTRATION IN CALIFORNIA 1 (2002), *available at* http://www.courts.ca.gov/documents/sipes_intro.pdf. The Council consists of 21 voting members[77] who are assisted by advisory members and Council staff.[77] The Council is meant to be responsive to the public as a whole, and not to any particular constituency.

[78] CAL. R. CT. 10.1(b) ("Article VI, section 6 of the California Constitution requires the council to improve the administration of justice by . . . [a]dopting rules for court administration and rules of practice and procedure that are not inconsistent with statute.").

[79] For instance, one internal committee is the Rules and Projects Committee, which "establishes and maintains a rule-making process that is understandable and accessible to justice system partners and the public." *Advisory Bodies*, CAL CTS., http://www.courts.ca.gov/advisorybodies.htm (last visited Feb. 5, 2019). These advisory bodies are governed by the California Rules of Court. CAL. R. CT. 10.30. *See also* Judicial Council Governance Policies, JUD. COUNCIL CAL. (Nov. 2017), http://www.courts.ca.gov/documents/JCGovernancePolicies.pdf (manuscript at 4).

Assessment Services."[80] In other words, for each county, a PAS is to generate risk "scores or levels," and a statewide panel appointed by the Council is to designate which PAS "scores or levels" are associated with high, medium, or low risk levels.[81] The choice of the risk threshold—which triggers the decision about how to treat an individual—is thus a global one.

Yet applying this global choice may result in different outcomes in different local jurisdictions. As Part III explores, applying the same high, medium, and low risk cutoffs in counties with different demographic distributions may produce different racial, gender, or socioeconomic effects, by county. The COMPAS debate about algorithmic unfairness arises in part from this point: if the baseline rate of arrest is different for white and black defendants in two jurisdictions, and the same high/medium/low risk categorization applies globally, then different proportions of individuals from each race will be affected in each jurisdiction.[82] Reserving any normative critiques of such a result, the practical upshot is the link between global and local authority. In a regime like SB 10 that relies on a centralized definition of risk levels, global policy choices set levels of risk at particular numerical points. These global decisions implement a particular technical understanding of fairness that controls local outcomes, potentially without accounting for local differences (as a technical matter). This policy outcome is the by-product of technical and policy constraints, as opposed to a consciously-pursued and explicitly stated definition of what is fair.[83] Without designing a statute to account for these global-local tensions and tradeoffs, we risk encoding these understandings implicitly, in ways that are opaque and may resist democratic accountability.

To make the real-world effects of risk assessment policies more transparent, policymakers and technologists should work together to parse each piece of the statute, in context, before it is enacted. In the case of SB 10, other portions of the statute go the other direction: they grant more authority to localities, with less precise global guidance. Specifically, SB 10 specifies a

---

[80] *See* §1320.25(b); 1324(e)(7) (directing the Judicial Council to "convene a panel of subject matter experts and judicial officers" that is to "designate 'low,' 'medium,' and 'high' risk levels based upon the scores or levels provided by the instrument for use by Pretrial Assessment Services"). This text is confusing as written because the PAS "risk score" may "include a numerical value or terms such as "high," "medium," or "low" risk," yet the high/medium/low "risk level" is also to be set by the Council, based on the PAS risk score. *See* § 1320.25. It nonetheless appears that the Council is responsible for convening the panel that designates the final risk levels.

[81] The statute does not clearly state whether the threshold set by the Council can be county-specific, or whether it must be uniform across the state. If it must be uniform across the state, this would imply a coordinated calibration of the models, which would stymie local validation of the instrument. If it can be county-specific, then there would be more local variation regarding how the statute treats individuals at a given quantitative risk level. For further discussion of this tension, see *infra* Parts III-IV.

[82] The COMPAS debate centers on the use of different base rates within a single county. As explored in detail *infra* Part III, an important, yet underexplored technical consideration involves the proper unit of analysis. This Essay considers *inter*-county differences, including how different county-by-county racial demographics further complicate (as a technical matter) the initial decision of how to set global and local authority (as a policy matter).

[83] Other risk assessment bills are more explicit on this point. For example, in contrast to SB 10, Idaho House Bill No. 118 provides an explicit definition of fairness: "'Free of bias' means that an algorithm has been formally tested and shown to predict successfully at the same rate for those in protected classes as those not in protected classes, and the rate of error is balanced as between protected classes and those not in protected classes." H.B. 118, 65th Sess., 1st Reg. Sess. (Idaho 2019), *available at* https://legislature.idaho.gov/wp-content/uploads/sessioninfo/2019/legislation/H0118.pdf. The Idaho bill thus requires predictive parity among protected classes. For a discussion of predictive parity, the measure adopted by the controversial COMPAS tool, see *supra* text accompanying note 7. The California statute does not contain such an explicit operationalization of what "fairness" requires. Whether the Idaho definition is normatively desirable is beyond the scope of this Essay, which focuses on how the design of a statute affects global and local development of and democratic control over a risk assessment instrument.

presumptive outcome for individuals that PAS assesses as "high-risk" and "low-risk," subject to judicial override as well as a host of exceptions enumerated in the text.[84] For individuals deemed high and low risk, there are global rules.[85] Where PAS concludes that an individual is medium-risk, however, it is to recommend their release or detention according to "standards set forth in the local rule of court."[86] SB 10 provides in subsequent text that each SC is to set these local rules of court "in consultation with Pretrial Assessment Services and other stakeholders." The statute additionally authorizes local rules that "expand the list of exclusions for persons assessed as medium risk that Pretrial Assessment Services is not permitted to release," so long as some medium risk individuals are still released.[87] Apart from this requirement, the local courts are constrained only by the requirement that local rules are "consistent" with the Council's global rules of court.[88]

Though the relevant Council rules are in stasis for the foreseeable future,[89] the draft versions do not offer much more concrete guidance to SCs. Draft Rule 4.40 provides, "[e]ach local rule must authorize release for as many arrested persons as possible, while reasonably assuring public safety and appearance in court as required."[90] Without more, however, these goals of "public safety" and "appearance in court" may not amply guide or constrain local actors. As Human Rights Watch warns in its comments on this proposed rule, "[t]his statement of purpose needs some specific regulations to make it meaningful."[91] Might there be reasons for local tailoring of these policy

---

[84] *See* § 1320.20 (high risk); § 1310 (low risk). As the Judicial Council's proposed rules explain:
"Prearraignment release of arrested persons will depend on their assessed risk level, determined by their score from the risk assessment tool and other information gathered from an investigation done by Pretrial Assessment Services, as follows:
- Low risk : Pretrial Assessment Services must release persons assessed as low risk prior to arraignment, on their own recognizance except for those persons arrested for misdemeanors or felonies who fall within the exclusions listed in section 1320.10(e). (Pen. Code, § 1320.10(b).)
- Medium risk: Pretrial Assessment Services has authority to release on own recognizance or supervised own recognizance, 9 or detain prearraignment, except for those persons subject to one of the exclusions listed in section 1320.10(e) or additional exclusions that may be included by a local court rule.10 (Pen. Code, § 1320.10(c).)
- High risk: Pretrial Assessment Services—and the court, if the court provides prearraignment review—is not authorized to release persons assessed as 'high risk.' Under sections 1320.10(e) and 1320.13(b), these persons must be held until arraignment when the court will make a release determination and set conditions of release, if applicable." *Criminal Procedure: Proper Use of Pretrial Risk Assessment Information; Review and Release Standards for Pretrial Assessment Services for Persons Assessed as Medium Risk*, CAL. CTS. https://www.courts.ca.gov/documents/SP18-23.pdf (manuscript at 2 (internal citations omitted)) [hereinafter *Proposed Rules 4.10 & 4.40*].

[85] *See id.*

[86] §1320.10.

[87] §1320.11(a). *See also Proposed Rules 4.10 & 4.40*, *supra* note 84 (manuscript at 13) ("If a court chooses to add to the list of exclusionary offenses or factors, the court must not adopt a rule that includes exclusions that effectively exclude all or nearly all persons assessed as medium risk from prearraignment release.").

[88] §1320.11(a).

[89] *See supra* note 3.

[90] *Proposed Rules 4.10 & 4.40*, *supra* note 84 (manuscript at 13).

[91] *Human Rights Watch Comments on California Judicial Council Bail Reform Rules*, HUMAN RIGHTS WATCH (Dec. 10, 2018, 9:00 AM), https://www.hrw.org/news/2018/12/10/human-rights-watch-comments-california-judicial-council-bail-reform-rules [hereinafter *Human Rights Watch Comments*]. *See also e.g.*, *Written Comments of the Electronic Frontier Foundation on Proposed Rules 4.10 and 4.40—Invitation to Comment #SP18-23* (Dec. 14, 2018), https://www.eff.org/files/2018/12/18/written_comments_of_the_electronic_frontier_foundation_on_proposed_rules_4.10_and_4.40_invitation_to_comment_sp18-23_december_14_2018.pdf (manuscript at 15) [hereinafter *EFF Comments*] ("Rule 4.40 is meant to prescribe the parameters of local rules . . . The proposed rule, however, provides no meaningful parameters on the

choices? Perhaps. Yet there are also institutional tradeoffs when global rules rely on local determinations to craft the policy. A framework that requires too much localization to craft the basic rules of the policy will not be globally consistent, and that lack of uniformity might itself be seen as unfair. Regardless of the equilibrium that is ultimately struck, when human life and liberty are at stake, these choices should be made intentionally, with an awareness of the tradeoffs. But the ways that risk assessment algorithms interact with systems of local and global discretion presently appear to be the inadvertent consequences of particular policy choices—not intentionally selected outcomes. With an emphasis on global-local tensions, Part III surveys some of the ways in which implementing SB 10 as a technical matter entails policy judgments that are not explicitly specified by the statute.

III. Algorithmic Operation in Context

Using SB 10 as a concrete example of a risk assessment statute, this Section considers how a statistically-driven risk assessment tool might operate in practice. The concrete narratives and associated modeling expose how even well-intended actors can produce unintended consequences when different kinds of demographic and algorithmic discrepancies interact with policy levers and zones of discretion at different jurisdictional levels (here, state and county). It focuses on how the statute interacts with three specific categories of technical issues—proxies, Simpson's paradox, and thresholding.

A. Proxies

As a policy matter, SB 10 and the available draft rules address fairness both globally and locally. Globally, the California Board of State and Community Corrections is to contract with independent third parties to assess the statute's effects, with an emphasis on "the impact of the act by race, ethnicity, gender, and income level."[92] Locally, the implementing court is to "consider any limitations of risk assessment tools in general, and any limitations of the particular risk assessment tool used by [PAS], including . . . Whether any scientific research has raised questions that the particular instrument unfairly classifies offenders based on race, ethnicity, gender, or income level."[93] Since the global requirements refer to independent audits, this Section emphasizes local fairness oversight. In concrete terms, what might local consideration of risk assessment limitations require, and what research might be relevant to determine whether there has been unfair classification?

Suppose that a risk assessment instrument eliminates any use of a sensitive attribute, such as race, in making predictions about whether a given individual will recidivate. As the growing fairness literature documents, this solution does not guarantee that the tool is free of bias based on that attribute. The problem is one of *proxies*,[94] a phenomenon wherein other valid features can be

---

use of pretrial risk assessment information, and fails to place any meaningful limitations on the discretion of Pretrial Assessment Services to detain individual defendants.").

[92] § 1320.3(a).

[93] *Proposed Rules 4.10 & 4.40*, *supra* note 84 (manuscript at 11–12).

[94] Here, a proxy variable is a feature that can be used to infer one or more of the protected attribute values of an individual. A well-known example proxy for race is zip code.

highly correlated with the protected attribute. Where a proxy for a protected attribute exists, decisions that incorporate a proxy variable may be biased even when the decision does not explicitly incorporate the protected attribute.[95] Consider a non-technical example: the problematic practice of "redlining" a neighborhood to infer racial demographics, thereby permitting racial discrimination without explicit consideration of race. Or at the individual level, a person's name can be a proxy for gender, such that making a decision about Aaron versus Erin could permit gender discrimination without explicit consideration of gender.

Simply forbidding the use of proxies for protected attributes is not a viable solution. Consider a feature like education. Information about an individual's education (e.g., highest degree, major, etc.) is correlated with gender,[96] such that education acts as a proxy for gender.[97] However, it is not feasible to dismiss it entirely as a factor in decision-making because it may in fact be relevant to the decision. For instance, the highest degree obtained or an individual's specialty area could be highly salient in hiring an individual. A proxy variable, in short, may include crucial information for making predictions. Instead of omitting them entirely, it may be more valuable to exploit them for prediction—while still limiting the bias due to their correlation with the protected attributes. In fact, several works have suggested that the only way to make a system truly fair, even from the effects of proxies, is to collect and take into account the protected attribute values at the learning and/or prediction stage.[98]

Several proxy-related issues arise under SB 10. Recall that the Council is to "[c]ompile and maintain a list of validated pretrial risk assessment tools,"[99] yet county-level SCs and the associated PAS are to ensure that the tool in use in a particular jurisdiction does not unfairly classify. To see

---

[95] *See* Eckhouse et al., *supra* note 25, at 15–16 ("[E]ven a determined effort to exclude proxies for race, class, or other marginalized categories [from an algorithm] is not likely to be successful. . . . [O]mitting race from the set of variables in the original data set does not mean race is not included in the analysis; it merely induces remaining variables that are correlated with both race and the outcome variable to behave as if they are, in part, proxies for race (citing Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 Cal.L. Rev. 721 (2016); Cynthia Dwork et al., *Fairness Through Awareness*, PROCEEDINGS OF 3RD INNOVATIONS IN THEORETICAL COMP. SCI. CONF., 214 (2012); Devin G. Pope and Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 Amer. Econ. J.: Econ. Pol., 206-231 (2011))).

[96] For a visualization of this data, which is provided by the US Census Bureau, see *Detailed Educational Attainment Sex Ratio*, STATISTICAL ATLAS (Sept. 4, 2018), https://statisticalatlas.com/state/California/Educational-Attainment#figure/detailed-educational-attainment-sex-ratio.

[97] Informally, correlation or dependence indicates whether a proxy variable can be used to predict one or more protected attributes. Approaches to quantify this relationship include Pearson correlation coefficient, which measures a linear relationship between variables, and mutual information, which measures probabilistic dependence between variables. Although not explored in detail here, proxies also raise practical difficulties. Determining both the correct statistical definition to quantify the proxy effect and whether a variable is an unacceptable proxy, either through numerical thresholds or qualitative assessment, demands substantial time and resources. These costly investments decrease the efficiency gains of a turn to algorithmic risk assessment.

[98] Some approaches include learning new representations of the non-protected attributes such that they still have predictive power while remaining independent of the protected attribute. These approaches generally require accounting for the protected attribute values at both the learning and prediction stage. At the learning stage, the restriction is placed on the data used to develop the tool. At the prediction stage, the restriction is on the features that the tool gets as input to make a prediction. *See* Richard Zemel et al., *Learning Fair Representations*, PROCEEDINGS OF THE 30TH INT'L CONF. ON MACHINE LEARNING 325 (2013); Michael Feldman et al., *Certifying and Removing Disparate Impact*, PROCEEDINGS OF THE 21ST ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 259 (2015). Moreover, one can make a learned tool fair by setting protected attribute specific threshold rules. *See* Samuel Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, PROCEEDINGS OF THE 21ST ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING, 797 (2017).

[99] § 1320.24(e)(1).

how this could play out, assume that the Council rules regarding validation—which were not completed before the statute was stayed—provide for local validation and adaptation of a tool that is on the approved list, in keeping with technical best practices.[100] Under this structure, each local SC is responsible for validating the risk assessment instrument used in its associated county, as a policy matter.

To see how this allocation of authority might operate, consider a hypothetical state of Idem, made up of Counties A, B, and C, in which SB 10's structure is applied. The counties are identically distributed for all observable measures.[101] They begin with the same risk assessment tool, selected from the list that Idem's statewide council has provided. Assume that the counties are able to tweak a pre-approved tool during the validation process.[102] They each proceed to validate and adapt the instrument, with an eye to ensuring it does not unfairly classify on the basis of the protected characteristic of race (as the statute requires).

	Size	H.S. Diploma
White	0.5	0.6
Black	0.5	0.4

Risk	No H.S.	H.S. Diploma	
White	0.825	0.25	0.48
Black	0.62	0.375	0.52
	0.7	0.3	

TABLE 1. LEFT: PROPORTIONATE SIZE AND HIGH SCHOOL COMPLETION RATE PER RACIAL GROUP. RIGHT: PROPORTION OF INDIVIDUALS IN STATE WHO POSE RISK, PER RACIAL GROUP AND EDUCATION LEVEL

Applying this tool and acting in good faith to apply the state's requirement about avoiding unfair classification, suppose County A and County B are each concerned that the use of any racial information at all is problematic.[103] They therefore decide to forbid risk assessment algorithms from

---

[100] *See* sources cited *supra* note 21 and accompanying text.

[101] Here, observable refers both to demographic characteristics and available information, such as past criminal history.

[102] Again, this point is ambiguous on the face of SB 10 without more explicit rules regarding validation, which are not yet published. This assumption is made to illustrate the broader claim that the very choice of where and how to permit validation (globally or locally) is a policy decision with technical implications that tend to be underexplored at the policy design stage.

[103] County A's concern could apply even if the officials are not acting with discriminatory purpose such that the U.S. Constitution's Equal Protection Clause would bar the consideration of race. As Sam Corbett-Davies and Sharad Goel explain, "the dominant legal doctrine of discrimination[] focuses on a decision maker's motivations. Specifically, equal protection law—as established by the U.S. Constitution's Fourteenth Amendment—prohibits government agents from acting with 'discriminatory purpose' (Washington v. Davis, [426 U.S. 229 (1976)). It bars policies undertaken with animus (i.e., it bars a form of taste-based discrimination, since acting with animus typically means sacrificing utility); but it allows for the limited use of protected attributes to further a compelling government interest (i.e., it allows a form of statistical discrimination)." *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, ARXIV 4 (Aug. 18, 2018), https://arxiv.org/abs/1808.00023. *Cf.* Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. (forthcoming 2019) (manuscript at 9) (articulating limitations of "Equal Protection jurisprudence in relation to algorithmic criminal justice" and offering that this jurisprudence "is not a coherent or morally acute metric").

For purposes of the above hypothetical, assume that there is no applicable *federal* statutory regime that supports a disparate impact analysis, that the state constitution does not outright bar discrimination or preferential treatment on the basis of a suspect classification like race or national origin, and that there is no explicit racial animus that could establish a federal constitutional violation under *Washington v. Davis* and its progeny. A state actor might nonetheless decide to avoid the use of a particular characteristic like race or gender. At least one recently-enacted statute presently takes a hardline approach of this sort

using race as a feature and also bar the consideration of proxies that might correlate with race, such as zip code. However, they make different choices about what constitutes unacceptable proxies. In the hypothetical distribution shown above, education is correlated with race, since white individuals are more likely to have a high school diploma than black individuals (60% vs. 40%). Based on this finding, County A bars the use of education level in the tool. On the other hand, County B does not require that the tool be completely independent of an individual's education level, on the grounds that it provides salient information about the risk that an individual poses. Lastly, County C permits the use of race as a feature—but the local PAS that administers the tool stipulates that it can only be used to correct for possible inadvertent bias from proxies. In other words, County C explicitly attempts to use race to achieve demographic parity, whereas the other counties bar the use of race or proxies for race. What happens when each county's tool is applied to a demographically identical individual?

Risk	No H.S.	H.S. Diploma	
White	0.85	0.27	0.5
Black	0.6	0.35	0.5

TABLE 2. RISK SCORES ASSIGNED BY TOOL IN COUNTY C, PER RACIAL GROUP AND EDUCATION LEVEL

Imagine the same individual who is assessed in each county. Recall that County A does not permit the use of race or education level as a feature. Thus, the individual is assigned a score of 0.5, which is the average risk of the entire population. This will be a blanket score assigned to defendants regardless of their race or education level. Although it can be considered fair as it makes no distinction between individuals, such a tool sacrifices accuracy as an assessment algorithm. On the other hand, even barring the explicit use of race, the tool in County B can infer an individual's race through their education level. As a result, a black defendant in County B with no high school diploma will be assigned a risk score of 0.7. However, County C uses race to correct for inadvertent bias from a proxy (here, education level). In fact, by considering both an individual's race and education level and thereby explicitly accounting for proxies, County C will achieve demographic parity in the manner illustrated in the above table. In County C, then, the same black individual with no high school diploma will be assigned a score of 0.6, lower than that in County B (0.7), but higher than that in County A (0.5). The upshot is that counties using the same instrument can each attempt to account for "any limitations of the particular risk assessment tool used by [PAS], including . . . [w]hether any scientific research has raised questions that the particular instrument unfairly classifies offenders based on race, ethnicity, gender, or income level," yet treat an identical individual differently because they interpret unfair classification differently in their treatment of proxies.

This result illustrates a global-local proxy tension. As a technical matter, a policy that permits each locality to determine how to handle proxies permits inconsistent decisions regarding treatment of proxies across counties, and thus allows for different outcomes for the same hypothetical individual, based on the county in which they are located. Yet as a policy matter, localized

---

and bans consideration of gender. *See* Ann Carrns, *In California, Gender Can No Longer Be Considered in Setting Car Insurance Rates*, NY TIMES (Jan. 18. 2019), https://www.nytimes.com/2019/01/18/your-money/car-insurance-gender-california.html (quoting California Insurance Department: "'Gender's relationship to risk of loss no longer appears to be substantial,' the department noted, saying the rationale for using it was 'suspect'").

discretion about how to validate a tool may be a good thing from the perspective of local self-determination. And the issue is even more complex because there may also be important technical reasons to permit local proxy determinations.

To see why, imagine that the global statute barred the use of particular protected attributes and their proxies. The trouble with such a global fiat is that the strength of the proxy (level of correlation) can vary in each county and in the state as a whole. In such a case, a county that does *not* exhibit a correlation between a protected attribute and a specified proxy would need to exclude variables that do not pose a huge proxy problem, thereby sacrificing accuracy. On the other hand, county-by-county determination of proxies is likely to be not only extremely resource-intensive, but also problematic insofar as the result is a patchwork of county-by-county policy calls that make global oversight challenging. Neither a global nor a local solution is perfect. But the bottom line is that we cannot hope to identify who should determine fairness at either the global or local level, with what parameters, if we do not begin with an awareness of ways in which local technical discretion gives substance to a value like fairness. And as we will see in the following two Sections, global policy requirements come with further costs.

B. Simpson's Paradox

Without a top-down, global understanding of what fairness requires, the interaction between local and global parsing of the data might independently be problematic. Specifically, applying SB 10's requirements, there is a risk of *Simpson's paradox*. Simpson's paradox refers to a statistical phenomenon in which a trend appears in several subgroups of a population, yet that same trend disappears or reverses when the groups are aggregated.[104] Consider, for instance, a group of men and women who work at a university that has two departments, Department A and Department B. Imagine a hypothetical algorithm that determines who receives a positive outcome, such as promotion in the department. In Department A, 25% of women receive a favorable decision, and no men receive that decision, such that there is evidence of systematic bias against these males. In Department B, 100% of females receive a favorable decision, and 75% of men do. Again, there is evidence that men are systematically less likely to receive the desirable result, even though the outcome is less skewed. Across the university as a whole, however, there is complete gender parity, such that global oversight alone would not reveal any evidence of unfairness:

	Female	Male
Dept. A	**0.25 (40)**	0.00 (20)
Dept. B	**1.00 (20)**	0.75 (40)
Combined	0.50 (60)	0.50 (60)

TABLE 3. PROPORTION OF POSITIVE DECISION (POPULATION SIZE)

This effect can also be easily found in real-life data. For instance, consider the following education level and ethnicity data.[105] The attainment rate of a four-year college degree in each of the seven California counties is higher among the white population than among the Asian population, but the trend is reversed when these counties are grouped together:

---

[104] Judea Pearl, *Understanding Simpson's Paradox*, 68 AMER. STATISTICIAN 8 (2014).

[105] This data covers 2011-2015 and relies on figures made available by the California Department of Public Health. *See Educational Attainment*, CHHS Open Data, https://data.chhs.ca.gov/dataset/educational-attainment (last visited Mar. 9, 2019).

	White	Asian
Fresno	**0.30 (219.3)**	0.28 (53.3)
Lassen	**0.16 (15.9)**	0.07 (0.3)
Marin	**0.62 (144.3)**	0.59 (11.4)
Monterey	**0.42 (104.6)**	0.36 (16.5)
San Mateo	**0.56 (236.3)**	0.55 (147.8)
Sutter	**0.22 (33.0)**	0.21 (9.2)
Tuolumne	**0.22 (34.2)**	0.11 (0.3)
Combined	0.44 (787.6)	**0.47 (238.8)**

TABLE 4. ATTAINMENT RATE OF A FOUR-YEAR COLLEGE DEGREE OR HIGHER (POPULATION SIZE)

In the context of algorithmic fairness, SB 10 would allow Simpson's paradox to occur if there is an opportunity for discrimination against a group at the local (county) level, yet the discrimination becomes undetectable or even reverses direction at the state level, or vice versa.

For example, an individual's education level is one factor that contemporary instruments often use.[106] How might this consideration play out in the SB 10 context? Suppose that at least some of the tools approved by the Council take education level into account. Suppose, further, that the seven counties shown above decide to adopt risk assessment instrument E, which considers education as a major feature. For the sake of argument, say that tool E assigns a low numerical risk score to individuals with a four-year college degree or higher, and that this assessment holds across each of the seven counties. Assume that this quantitative risk score falls into the "low risk" qualitative category for the state. In other words, across the state, tool E's assessment leads to the conclusion that individuals who have completed four or more years of higher education are, collectively, a low risk group.

Such a scenario could easily result in differential risk assessments for members of different demographic groups. Imagine, first, that an independent statewide auditor wishes to confirm that the tool is not unfairly classifying based on ethnicity. If such an entity validates the tool using aggregate data alone, as it is likely to do given the cost and difficulty of validating with reference to each individual county, then the tool could be rejected for discriminating against the white population, on the ground that the 0.03 difference between the respective white and Asian assignment rates to the low-risk group is unacceptable. Alternatively, it could be accepted on the grounds that the 0.03 difference between the white and Asian population in being classified as low risk is negligible. Each of these conclusions is a normatively-laden policy determination, embedded in the auditing process, that merits further independent discussion by scholars and policymakers. Assuming that a consensus on this matter is possible, the present point is that Simpson's paradox remains at the county level: a tool deemed globally fair would in fact be biased against Asians in every one of the seven counties. Returning to the hypothetical example of a tool

---

[106] For example, COMPAS, the tool that was used in the *Loomis v. Wisconsin* case, considers education level, *see* PAMELA CASEY ET AL., OFFENDER RISK & NEEDS ASSESSMENT INSTRUMENTS: A PRIMER FOR COURTS, NAT'L CTR. STATE CTS. (2014) (manuscript at A-21), and the Federal Pretrial Risk Assessment (PTRA) uses highest educational attainment as one of the features of its logistic regression model, *see* Christopher T. Lowenkamp & Jay Whetzel, *The Development of an Actuarial Risk Assessment Instrument for U.S. Pretrial Services*, 73 FED. PROBATION (2009), https://www.uscourts.gov/sites/default/files/73_2_3_0.pdf. Several counties in Maryland also rely on locally-developed tools that take education level into account. *See* Angela Roberts & Nora Eckert, *As Maryland Courts Meld Artificial Intelligence into Bail Decisions, Concerns Follow*, CAP. NEWS SERV. (Dec. 21, 2018), https://cnsmaryland.org/interactives/spring-2018/plea-bargain/pretrial-riskscore.html. *Cf.* Jessica Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 81 (2017) (discussing factors, including education level, used in existing tools).

for which achieving a four-year degree mediates the quantitative risk assessment level and results in a low risk label, this intra-county finding would mean that Asians receive comparatively higher risk scores than whites within each of the counties. Furthermore, such discrimination can be much more significant than the 0.03 difference observed at the state level. As the above chart illustrates, in Tuolumne for instance, white defendants are twice as likely to be in the low-risk group than Asian defendants. Accordingly, different local and global parsing of the same data set can make even the same outcomes seem fair or unfair, depending on which vantage point is adopted.

C. Thresholding

Specific policy provisions within SB 10 also carry unrecognized technical—and human—consequences. Recall that the statute requires the global Council to designate which PAS "scores or levels" are associated with high, medium, or low risk levels.[107] In operational terms, this means that each individual is first assigned a score by a risk assessment tool. Then, the individual is categorized into a risk group by thresholding their score: high risk if the score is above a certain value, low risk if it is below another value, and so on. It may be natural to assume that for any scores generated by a fair system, however fairness is defined, a given threshold will still guarantee fairness. However, the practice of thresholding can introduce a new set of challenges.[108] First, as the fairness literature has emphasized, there are several mathematical notions of fairness, each of which may require a different set of thresholds.[109] For example, many notions of fairness aim to achieve a balance of classifier performance metrics, such as positive predictive value, false negative rate, and false positive rate,[110] between different protected groups. The technical fairness literature has shown that no single threshold rule can satisfy all three of these fairness definitions except in an unlikely assumption;[111] thus, choosing a threshold to enforce one notion of fairness would violate another.

This Essay emphasizes a less-recognized second point: even with an agreed-upon definition of fairness, enforcing it among subpopulations—such as different ethnic groups—may be impossible without setting a different threshold for each group.

---

[107] Despite some textual ambiguity, the choice of the risk threshold—which triggers the decision about how to treat an individual—is up to the Council. *See supra* note 81. Again, it is not clear on the text alone whether the threshold set by the Council can be county-specific, or whether it must be uniform across the state. *Id.* The present discussion assumes that SB 10 provides for levels that are uniform across the state.

[108] For an interactive visualization of thresholding challenges, see Martin Wattenberg et al., *Attacking Discrimination with Smarter Machine Learning*, GOOGLE RESEARCH, http://research.google.com/bigpicture/attacking-discrimination-in-ml/ (last visited Mar. 9, 2019).

[109] *See* Narayanan, *Tutorial: 21 Fairness Definitions and Their Politics*, *supra* note 7.

[110] In the context of pretrial risk assessment, positive predictive value refers to the proportion of high-risk individuals who indeed recidivate or fail to appear in court; false negative rate is the proportion of people who would recidivate or fail to appear but were classified as low-risk; and false positive rate is the proportion of people who would appear in court but were classified as high-risk. *See also supra* text accompanying note 7 and sources cited therein..

[111] The exception is when the base rate (*e.g.*, the probability of recidivism or failure to appear) is equal among the protected groups. *See* Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study Of Bias In Recidivism Prediction Instruments*, ARXIV (Feb. 28, 2017), https://arxiv.org/abs/1703.00056. Interestingly, Northpointe cited the lack of equal base rates in the observed population as a defense of its risk assessment tool during the COMPAS controversy. *See supra* text accompanying notes 52–55.

FIGURE 1. SCORE DISTRIBUTION OF GREEN AND BLUE GROUPS—DIFFERENT MEAN
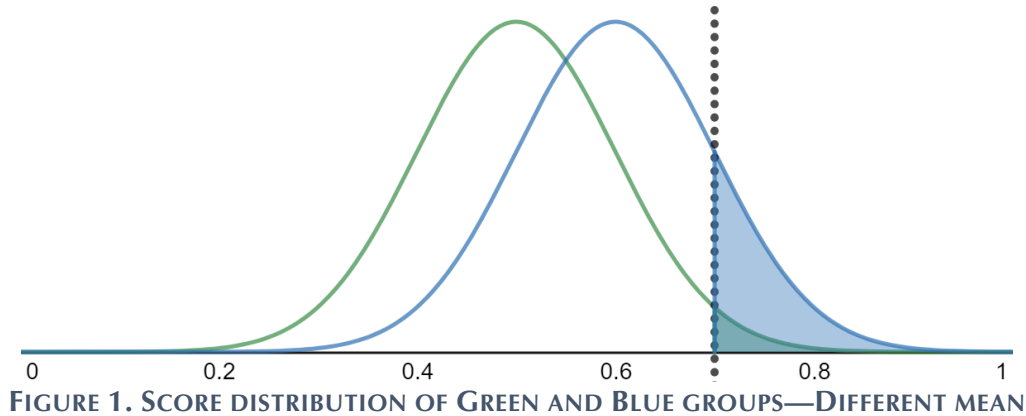
This thresholding consideration emerges in a risk assessment policy that applies a global risk threshold across local jurisdictions—including application of the same threshold to subpopulations in a given jurisdiction.[112] Suppose a county consists of two ethnic groups, Green and Blue, and adopts a risk assessment tool whose scores for the Green and Blue groups are each normally distributed with mean 0.5 and 0.6, respectively, with standard deviation 0.1. Imagine, further, that this tool has reached some agreed-upon balance of fairness and accuracy (however defined). Thresholding can still introduce challenges due to distributional discrepancy within the jurisdiction. Recall that, under SB 10, risk refers to the "likelihood that a person will not appear in court as required or the likelihood that a person will commit a new crime if the person is released before adjudication of his or her current criminal offense."[113] Applying this understanding, suppose that the global Council decides to categorize individuals with risk higher than 0.7 as the high risk group, indicated by the highlighted areas in above figure.

How might the state's global choice affect a given county, and the individuals within it to whom the tool is applied? Applying the scenario described above, among the Green individuals who would neither recidivate nor miss the court date, roughly 1% will be classified as high risk. On the other hand, this number is significantly larger—10%—for the Blue group. That is, a Blue individual who poses no risk is ten times more likely to be classified high-risk than had they been in the Green group. Formally, the false positive rates of Green and Blue groups do not match, and the categorization would be considered unfair under this notion of fairness. In fact, in order to achieve equal false positive rates using a single threshold, we need either to classify everyone as high risk (use a threshold very close to 0) or no one as high risk (threshold close to 1).

Nevertheless, we could still achieve a fair thresholding rule by setting a different threshold for each ethnic group. For instance, thresholding the Green individuals at 0.65 and Blue at 0.74 would achieve fairness with equal false positive rates of 4%. In policy terms, this would require more particularized, variable thresholding *within* a given population. But there is a tradeoff here between global uniformity and localized fairness at the level of subpopulations. The bottom line: even if we achieve fairness of risk scores, using any given definition of fairness, this result does not

---

[112] If each sub-jurisdiction sets its own risk threshold levels (e.g., determines what quantitative scores are associated with high, medium, and low risk categories), then the understanding of "risk" would be localized, without global consensus of the sort law typically demands.
[113] § 1320.7(h).

guarantee fairness of the threshold risk groups if a single threshold was used for different subgroups, such as different racial or ethnic groups.

Furthermore, this sort of thresholding problem can persist even when the base distributions of different subgroups are identical, due to uncertainty inherent in any statistically-derived risk assessment instrument. Consider a different county with Green and Blue ethnic groups. This county has identically distributed risk scores centered around 0.5. Suppose an algorithmic scoring system outputs a risk score with more variance for the Green group than for the Blue group. More precisely, the predicted risk scores have standard deviation 0.2 for Green and 0.05 for Blue, as shown in the following figure.
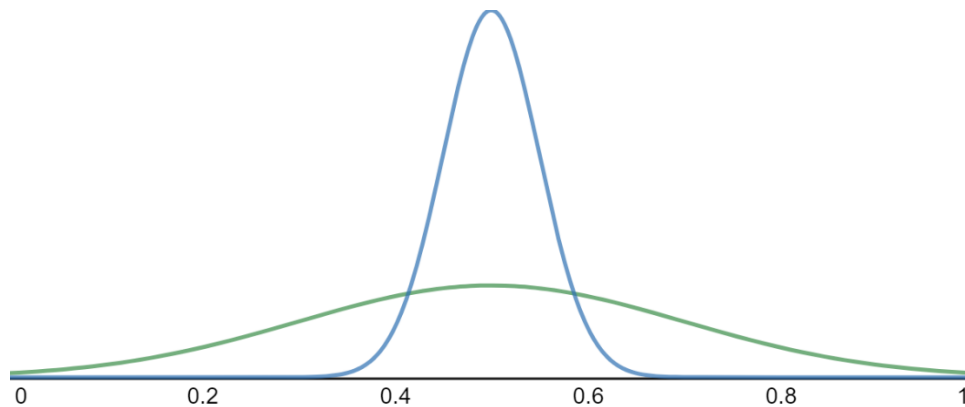
FIGURE 2. SCORE DISTRIBUTION OF GREEN AND BLUE GROUPS—DIFFERENT VARIANCE

Such a phenomenon can often occur when one group, in this case Green, makes up a greater proportion of the population. Because algorithmic instruments are built to optimize a measure (such as accuracy) for the overall population, they will perform better for a group that makes up a greater proportion of the whole. Put differently, the tool will have a better predictive power for the majority group. Such a tool can be considered fair in the sense that the scores are equally well-calibrated for both groups in the population as a whole; for example, any individual assigned a score 0.8 will indeed have 0.8 probability of committing a new crime or failing to appear in court. But the combination of the tool's quantitative scoring and the thresholding decision can still be considered unfair for members of non-majority subpopulations.

Again, applying the parameters of a statute like SB 10, suppose that a global body sets a threshold of 0.7 to classify an individual as high risk, regardless of their group membership in Green or Blue. In such a situation, about 81% of Green individuals in the resulting high-risk group would have recidivated or failed to show up, whereas this proportion drops to 71% for the Blue group. This is an example of a classifier failing to satisfy fairness as defined by predictive parity[114] (the balance in positive predictive value between ethnic groups). On the other hand, setting a threshold for each group independently could, in theory, achieve a more subgroup-sensitive understanding of fairness. But this result would require far more localization of the overarching policy categories.

---

[114] Predictive parity is sometimes referred to as calibration in the technical fairness literature.

Thresholding of high, medium, and low levels can thus lead to unfair risk categorization in at least two ways. One, there may be unfairness due to distributional differences between subpopulations (e.g. ethnic, gender, or age groups). Two, there may be unfairness due to the inherent uncertainty in statistical risk assessment tools. Policy guidance must, accordingly, be clear not only about the need to eliminate "bias," or "unfair" classification, but also about which level(s) of authority are to do so, with what amount of specificity in classifying a population.

In the case of a statute that sets global thresholding standards across the entire jurisdiction, as SB 10 appears to do, these sorts of thresholding issues become especially stark. In addition to underlying technical and normative questions about the "right" thresholding practices, the way that the statute or regulation allocates authority becomes critical. In particular, any global delineation might be difficult or even impossible to correct when the entities tasked with correcting unfairness are local. For instance, SB 10's draft guidance requires local courts to consider whether "any scientific research has raised questions that the particular instrument unfairly classifies offenders based on race, ethnicity, gender, or income level." Scientific research might indeed raise questions—but whether the questions matter in fairness determinations cannot be answered in the abstract. Rather, the very definition of "unfairly classify[ing]" is bound up in the antecedent global choices about risk thresholds, such that this sort of global/local allocation risks asking local actors to account for choices over which they have no input or control.

This issue can, moreover, further be compounded with demographic discrepancies among counties and/or between counties and the state. For example, versions of Simpson's paradox mean that a threshold (or a set of thresholds) chosen to be fair at the state level may not achieve fairness at the level of individual counties. Recall how, in the first Green/Blue group example, the original threshold of 0.7 was adjusted down for the Green group and up for the Blue group in order to make the categorization fair (in the sense of equal false positive rates). Suppose this choice was made at the global (state) level. If Simpson's paradox is present and the distributional difference in each local (county) reverses the direction (i.e., the Green group in fact has a higher mean risk score at the level of the locality), then this pair of thresholds set by the state would be making the risk categorization less fair at the global (state) level. Conversely, suppose once more that every county uses the same risk threshold in an attempt to be fair across the state. This tack is likely to run into different technical issues: an attempt to threshold such that the resulting risk categorization is fair in every county may not exist unless the thresholding itself is minimal, such as, for instance, a system that assigns everyone to a single risk group. In other words, an attempt to be fair at the global level may end up being unfair at the local level, yet an attempt to be fair at the local level may be technically impossible. This global-local tension is a hidden policy consequence of the technical choices required to build risk assessment instruments.

* * *

The cumulative upshot of these sorts of technical considerations is that risk assessment tools require complex webs of policy and technical choices, globally and locally. Maximizing technical objectives may require policy tradeoffs, and vice versa. Moreover, not all policy interventions are created equal when it comes to technical consequences; rather, granting authority to a local versus a global decision-maker interacts with technical risks and affordances. A failure to begin with this ground-level awareness is tantamount to creating black box policy regimes that turn out to be Pandora's Boxes if we try to open them down the line. The following Part thus builds from the

specific tradeoffs observed in a statute with terms like SB 10 and begins to distill more generally applicable principals for the design of risk assessment statutes and regulations.

IV. Paths Forward

Abstracting away from SB 10's particulars, every risk assessment algorithm inevitably entails local and global allocations of authority along two related axes. One, which entity, at which level, is responsible for crafting the relevant procedures and rules or standards. This is the more traditional *policy* prong of a statute or regulation. Two, which entity, at which level, is responsible for developing, testing, and applying the instrument itself, potentially subject to local or global constraints or guidance. This is the more *technical* prong of a statute or regulation.

In practice, moreover, the picture is even more complicated because risk assessment does not allow such a crisp bifurcation of technical and policy choices.[115] Technical choices must account for local conditions to avoid unfair results, yet law's commitment to global first principles cuts against too much tailoring by jurisdiction.

This tension is especially stark for a multi-level intervention like SB 10 because the Council's proposed rules of court require each SC and its associated PAS to ensure that the tool is "accurate," to assess whether it has been appropriately validated, and to consider whether there has been unfair classification.[116] Yet neither "fairness" nor the normatively proper tradeoffs between fairness and other values, like accuracy, are self-defining. Nor is there further delineation in either the statutory text or its legislative history to clarify what it means for, say, an instrument to "unfairly classif[y]" based on a sensitive characteristic. The substance of these normative requirements, accordingly, will be defined locally. And given this inevitable tension between local tailoring and global commitments, clear oversight of the system itself is all the more critical to ensure that the system itself remains accountable. Too many *layers of discretion* at both the policy and technical levels risk creating a policy black box in which implementing an algorithm channels authority in unanticipated directions, potentially without adequate democratic responsibility for the ways in which the algorithm affects actual human lives.

In the face of such complexity, we advocate simplicity. The SB 10 example suggests that risk assessment statutes or regulations will run into trouble where they create too many layers of discretion. There are two specific issues. One, if there are zones of ambiguous or even conflicting control (such as, for instance, a top-down, global definition of high, medium, and low risk and a locally validated tool), then there are likely to be disparate effects across counties that undermine any effort to create policy outcomes that are globally consistent. Two, the very process of technical validation demands local determination, and an attempt to too-strictly control the tool's development and implementation top-down will undermine any effort to create risk assessment

---

[115] *Cf.* Alicia Solow-Niederman, *Administering Artificial Intelligence* ("Algorithmic and programming decisions structure human behavior. These choices are in fact policy decisions that function at the most essential levels of democratic governance and public interests. Put simply: AI development is an especially stark example of how private coding choices are governance choices.") (manuscript at 48) (manuscript on file with author).

[116] *Proposed Rules 4.10 & 4.40, supra* note 84 (manuscript at 11–12).

instruments that are locally accurate and unbiased. It is beyond the scope of this analysis to endorse more global or local control; such theoretical development of what to weigh in crafting a system of algorithmic governance and how to strike the right balance of global and local when it comes to both technical and policy choices awaits future research.[117] Nonetheless, simplicity counsels in favor of several preliminary lessons, with associated implications for SB 10.

Less Can be More. Tools that attempt to introduce more factors might increase accuracy, so long as the information is managed properly. Yet they might also contain more opportunities for issues, like Simpson's paradox hidden discrimination, that can elude oversight systems, particularly systems that operate at a global level and must account for many localities. For SB 10, the tools ultimately adopted, if any, should use the minimum number of factors that avoids problematic thresholding variance.

Timing Matters. Policy requirements should take care in prescribing *when* global oversight is helpful, and what each global and local actor, respectively, is permitted to do at different stages of the tool development and deployment process. For example, if there is a list of globally approved tools that have been validated, as is the case for SB 10, may a locality undertake further validation to respond to a demographic change or a local policy, such as bail reform measures, that affect the likelihood of nonappearance? Risk assessment tools will make stale predictions if they are trained on historical data that does not account for more recent bail reforms.[118] To ensure that localities can update their instruments to reflect changing conditions on the ground, risk assessment statutes should both require ex post auditing of locally validated tools and be careful in calling for pre-approval of tools that are removed from a particular local context.

Audit with Attention to Local and Global Detail. Several of the technical challenges—most notably Simpson's paradox—that arise in SB 10 occur because of a lack of adequate attention to local data. When it comes to auditing the way that a tool classifies individuals, aggregated data is not enough. Audits must take into account localized data, not merely aggregate data. Though doing so is more cost- and time-intensive, this allocation of resources must be made before deploying the tool, as part of the initial cost-benefit analysis of a turn to risk assessment algorithms.

Mixed Zones are a Mixed Blessing. Policy provisions that demand both local and global control may seem like a helpful compromise—yet if they are not administered carefully, then they can complicate oversight of and public accountability for risk assessment instruments. Attempts to combine local and global oversight—as seen, for instance, in SB 10's requirement that local courts assess the accuracy and discriminatory potential of tools selected from a Council approved list—introduce a number of wrinkles.

Take validation, for instance. If a global body like the Council truly validates a tool, then it is not clear how a locality could adapt it to meet technical best practices and still permit global confidence in the tool. On the other hand, if the locality validates the tool, then substantial resources will be required for the global body to be certain that it meets its validation requirements—or there will be no meaningful oversight.

---

[117] *See supra* note 14 and accompanying text.

[118] *See* Koepke & Robinson, *supra* note 21, at 24–38 (warning against "zombie predictions" that rely on stale data).

The most auspicious way to manage these global-local tensions is to approach mixed zones with caution. Caution in crafting a policy means proceeding with an ex ante awareness of which level(s) must participate at a given stage in the process, as a technical matter, and how these considerations align with the allocation of decision-making authority, as a policy matter—and thinking carefully about how to grant decision-making and oversight responsibility, particularly if a proposed policy requires both local and global participation.

A Dash of Discretion. When used sparingly, allocating additional discretion within the statute might at times solve particularly thorny technical issues. For instance, in thresholding, SB 10 appears to provide that the same high/medium/low risk threshold set shall be applied to already-developed assessment tools. However, as Part III describes, this situation could lead to an impasse where the risk group assignment cannot be made fair without setting different thresholds for different protected subgroups or altering the tools with respect to the given threshold set. In the former instance, the global entity that sets the threshold would need additional discretion to set different qualitative risk levels associated with different subgroups (however identified). In the latter instance, localities would need additional discretion to develop and validate tools that meet the subgroup-independent thresholds in the context of that locality. The bottom line is that there must be additional tailoring if the global thresholds are to be considered fair in particular counties, whether the discretion to tailor as required is allocated globally or locally.

Define "Fairness" and Specify Who Decides. Even if no single definition of fairness is likely to be without controversy, risk assessment statutes should say what they intend as a *technical* matter when it comes to such a critically contested term. When these points are unspecified, they still must be made, but the choices will tend to be implicit—as a matter of technical development— without upfront consideration, adequate opportunity for public debate, or ongoing accountability for the decision. By, first, clearly defining whether a global or local entity is responsible for arriving at what is "fair," and, second, designing policies that designate what fairness means as a technical matter, we can better begin to grapple with the underlying normative implications of the statutory text. A critical matter for future research is whether, as a normative matter, fairness must be defined locally– lest the top-down imposition infringe on community values – or globally – lest too much tailoring to community norms contravene non-negotiable first principles.[119]

Conclusion

Developing and deploying risk assessment algorithms without considering how they will fit within new and existing institutions, norms, and preexisting technical and policy constraints is a mistake. The example of SB 10 highlights how risk assessment tools are not instruments that operate in isolation; rather, they are developed and deployed within legal institutions and require input from global and local decisionmakers. Control of these instruments, in turn, requires keener attention to the design of risk assessment policies, and specifically to who is granted authority and discretion over the tools. When a particular statute or regulation empowers an actor at the global level to develop a list of approved tools, for instance, how does this choice interact with the technical needs of local actors or cabin local policy discretion? Conversely, if a statute or regulation requires

---

[119] We reserve for future research whether the political and legal arguments for federalist systems, and the perceived benefits of policy localization, apply with equal force in this context or are overcome by unique technical considerations.

local validation, what limitations does this place on global oversight of the tool? This Essay encourages policymakers and technologists to ask these and related questions, by design.

Initial statutory and regulatory decisions should thus be made with a keen awareness of local-global tradeoffs, technical limitations, non-negotiable policy objectives, and underlying normative principles. A failure to grapple with these questions will not erase them. Where there are too many layers of discretion and too many local-global tensions, we would be ill-advised to rely on algorithmic risk assessment instruments as criminal *justice* tools.