ELECTRONIC FRONTIER FOUNDATION

December 14, 2018

Judicial Council of California
Criminal Law Advisory Committee
455 Golden Gate Avenue
San Francisco, CA 94102-3688
By email attachment to:
invitations@jud.ca.gov
Eve Hershcopf, Eve.Hershcopf@jud.ca.gov
Kara Partow, Kara.Partow@jud.ca.gov

Re: Written Comments of the Electronic Frontier Foundation on Proposed Rules 4.10 and 4.40—Invitation to Comment #SP18-23

Dear Ms. Hershcopf and Ms. Partow:

The Electronic Frontier Foundation (EFF) submits the following comments on the Judicial Council of California's Invitation to Comment SP18-23, relating to the proper use of pretrial risk assessment information (Proposed Cal. Rule of Court 4.10) and review and release standards for persons assessed as medium risk (Proposed Cal. Rule of Court 4.40).

EFF is a member-supported, nonprofit, public interest organization based in San Francisco dedicated to protecting privacy and civil liberties in the digital age. Founded in 1990, EFF represents tens of thousands of dues-paying members. EFF works to encourage and challenge companies, government, and courts to safeguard civil liberties as new technology changes society. EFF is particularly concerned about the use of actuarial pretrial risk assessment tools in the criminal justice system. These tools are intended to reduce bias and discrimination, but when not carefully evaluated and validated, they can perpetuate the same sort of discriminatory outcomes as existing systems that rely on human judgement. They can also result in new, unexpected errors.

EFF advocates for specific criteria that pretrial risk assessment tools must satisfy to ensure transparency, accountability, and due process.[1] Risk assessment tools should not be introduced into any courtroom unless these criteria are satisfied. EFF has also

[1] *See* Hayley Tsukayama and Jamie Williams, "If A Pre-Trial Risk Assessment Tool Does Not Satisfy These Criteria, It Needs to Stay Out of the Courtroom," *Deeplinks* (Nov. 6, 2018), https://www.eff.org/deeplinks/2018/11/if-pre-trial-risk-assessment-tool-does-not-satisfy-these-criteria-it-needs-stay.

joined more than a hundred advocacy groups to urge jurisdictions in California and across the country already using these algorithmic tools to stop until they have considered the many risks and consequences of their use.[2]

We urge the Judicial Council to modify Proposed Rules 4.10 and 4.40 to place meaningful restrictions on the use of actuarial pretrial risk assessment tools, and the information generated by these tools, in California courts. While it is clear from both proposed rules that the Judicial Council recognizes that pretrial risk assessment tools are not without their flaws, and that care must be taken when integrating complex algorithmic models into judicial decision making, neither proposed rule places sufficient limitations on the use of risk assessment information by courts.

I. Rule 4.10: The Proper Use of Risk Assessment Information

Rule 4.10 is meant to prescribe the proper use of pretrial risk assessment information by courts—which includes both judges and Pretrial Assessment Services staff, who are officers of the court. *See* Cal. Pen. Code §§ 1320.24(a)(1), 1320.7(g). The proposed rule, however, focuses almost entirely on factors that courts must consider *alongside* a pretrial risk assessment score. It places no limitation on *how* or *under what conditions* pretrial risk assessment information may be used in the first instance, and fails to protect defendants' due process rights. It also fails to ensure that the use of pretrial risk assessment information will "[a]ddress any biases in pretrial release and detention decision-making," as the proposal says is intended. *See* Proposed Rule 4.10(a)(2)(C).

A. Rule 4.10 Must Require Courts to Separately Consider the Specific Risks Being Assessed.

The risk assessment instruments available today are actuarial tools based on statistics informed by human insight. Specifically, they are formulas that converts information about the person being considered (gathered, *e.g.*, from their personal attributes, criminal history, or a behavioral interview) into a score representing their risk of *something* (*e.g.*, violent recidivism, or failing to appear for a scheduled court date).

Rule 4.10 must mandate that risk assessment tools clearly distinguish between the types of risks being assessed, and define when an individual is assigned a risk score in one category and not the other, so that courts can effectively identify appropriate

[2] *See* Open Letter, The Use of Pretrial Risk Assessments: A Shared Statement of Civil Rights Concerns, http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf.

conditions to place on defendants for release. The risk of failing to appear is separate and distinct from the risk of committing an offense that presents a threat to public safety. And the risk of intentional flight itself may be importantly different from other forms of failure to appear.

The quantification of each of these risks "require[s] different assessments, based on different factors" and each should be "separately considered and weighed in accordance with applicable legal standards in the context of a given pretrial decision."[3] Indeed, "[a] high risk of failure to appear in court due to mental health issues is not the same as a high risk that a defendant will commit a violent crime while awaiting trial."[4] It is improper for courts to rely on pretrial risk assessment tools that conflate separate categories of risk or that fail to provide an explanation for how they arrived at their result.

Reports generated by risk assessment tools must also reflect that the tool is fit for its underlying purpose. Because the risks assessed pursuant to SB 10 are the "risk of failure to appear in court as required" and "*risk to public safety* due to *commission* of a new criminal offense *while released on the current criminal offense*,"[5] the risk assessment reports must differentiate between failure-to-appear, non-violent recidivism, and violent recidivism, as well as between mere rearrest and actual convictions, and be precise about the specific period for which the risk assessment applies. The objective of the tool is not to assess how likely a person is show up for a court date at any point in the future, but only during the period released for the specific charges at hand. Likewise, the objective of the risk assessment is not to determine whether the defendant is likely to be convicted of *or arrested* for any other alleged offense in the future; it is to determine how likely a defendant is likely to *commit a violent offense in the pretrial period* that justifies pretrial detention. Rearrest is not a good proxy for future criminality; it is well documented that different demographic groups are stopped, searched, arrested, and charged at very different rates across the United States.[6]

[3] Chelsea Barabas, Christopher T. Bavitz, Ryan H. Budish, Karthik Dinakar, Cynthia, Dwork, et al., An Open Letter to the Members of the Massachusetts Legislature Regarding the Adoption of Actuarial Risk Assessment Tools in the Criminal Justice System, 5 (Nov. 9, 2017), available at https://dash.harvard.edu/handle/1/34372582 (hereinafter "Barabas, *et. al*, Open Letter") (signed by signed by Harvard and MIT-based faculty, staff, and researchers).

[4] *Id*.

[5] *See* Cal. Pen. Code § 1320.7(b)–(d).

[6] *See, e.g.*, Marc Mauer, *Addressing Racial Disparities in Incarceration*, Supplement to 91 Prison Journal (2011); Joshua Rovner, The Sentencing Project, Policy Brief: Disproportionate Minority Contact in the Juvenile Justice System (2014),

B. Rule 4.10 Must Set Forth a Policy-Based Decision-Making Framework for How a Predicted Probability of the Risks Being Assessed is Quantified into a Risk Score—and How that Risk Score Should be Used and Interpreted by Courts.

Proper use of risk assessment tools by courts requires proper human insight and proper judgment about what the numbers mean and how they should be used. Risk assessment tools employ statistical methods to produce risk scores, and judges and Pretrial Assessment Services will consider those numerical scores as one factor in their pretrial decision-making process. It is vital that judges, Pretrial Assessment Services staff, and any other stakeholders in the pretrial pipeline "be trained to accurately interpret and understand risk assessment tools and the meaning (and limitations) of the risk assessment scores they produce."[7] This is important given *automation bias*, which refers to the tendency to view machines as objective and inherently trustworthy—even though they are not.[8] In addition to "training the staff who administer it and the judges and correctional/supervision staff who use its results," the proper use of risk and needs assessment requires "administering the right risk and needs assessments at the right time in the criminal justice process and accurately communicating those assessment results; adhering to a formal quality-assurance process; evaluating the data that come from risk and needs assessment to ensure that the assessment is working (*i.e.*, predicting recidivism accurately); and revalidating as necessary."[9]

It also requires the principled and fair application of risk assessment scores across the state. As one study uncovered, after analyzing more than one million criminal cases in Kentucky between 2009 and 2016 to determine how risk assessments affected pretrial outcomes, judges in rural and non-rural areas adhered to the risk assessment

http://sentencingproject.org/doc/publications/jj_Disproportionate%20Minority%20Contact.pdf.

[7] *See supra*, note 3, Barabas, et. al, Open Letter, at 4.

[8] Wikipedia, "Automation Bias" (last updated Dec. 3, 2018), https://en.wikipedia.org/wiki/Automation_bias ("Automation bias is the propensity for humans to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct").

[9] *See* Council of State Governments Justice Center, Risk and Needs Assessment and Race in the Criminal Justice System, Justice Center (May 31, 2016), https://csgjusticecenter.org/reentry/posts/risk-and-needs-assessment-and-race-in-the-criminal-justice-system/ (hereinafter "Risk and Needs Assessment and Race").

recommendations differentially, exacerbating racial inequalities.[10] In another study in Canada, researchers describe "criteria tinkering," an even more egregious example wherein court officers manipulate input values to obtain the score they think is correct for a particular defendant.[11]

As Harvard and MIT-affiliated faculty wrote in a letter to the Massachusetts Legislature regarding a proposed pretrial risk assessment bill, "[t]he classification of a risk category applicable to a particular criminal defendant with respect to a given risk score (*e.g*., high risk, medium risk, or low risk) is a matter of policy, not math"; tying loaded terms like "high risk" to scores generated by risk assessment tools influences both "decision-making by prosecutors, defendants, and judges in a pretrial setting (who may place undue emphasis on numerical scores generated by computers)" and "public perception of the specific outcomes of [risk assessment] tools."[12]

It is therefore essential that the Judicial Counsel clarify how risk scores should be generated and what they purport to predict, and set thresholds for low, medium, and high risk based on specific policy objectives of the state. When determining these policy objectives, California should be asking and answering the following question: "What trade-offs should we make to ensure justice and lower the massive social costs of incarceration?"[13]

Rule 4.10 must set forth a *policy-based decision-making framework* to guide interpretation of risk assessment predictions by all actors in the justice system across the state. This framework must be regularly updated to reflect ongoing research about what specific conditions have been empirically tested and proven to lower specific types of risk.

[10] Megan Stevenson, *Assessing Risk Assessment in Action*, 103 Minn. L. Rev., at 5, 43–44 (forthcoming, 2018), available at https://ssrn.com/abstract=3016088.

[11] *See* Hannah-Moffat, Kelly, Paula Maurutto, and Sarah Turnbul, *Negotiated risk: Actuarial illusions and discretion in probatio*n, Canadian Journal of Law & Society/La Revue Canadienne Droit et Société 24, no. 3, at 394 (2009), available at https://doi.org/10.1017/S0829320100010097 (noting that, in practice, practitioners "arrive at scores by incorporating their own experiences and their clinical knowledge of offenders"); Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, Justice Quarterly, at 16 (June 26, 2012), http://dx.doi.org/10.1080/07418825.2012.682603 (defining "criteria tinkering" as "adjusting the rating of individual items when filling out the forms").

[12] *See supra*, note 3, Barabas, *et. al*, Open Letter, at 4.

[13] Matthias Spielkamp, Inspecting Algorithms for Bias, MIT Technology Review (June 12, 2017), https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/.

Rule 4.10 must likewise mandate that courts, when consider risk assessment scores, consider the processes by which a predicted probability of failure to appear, or a predicted probability of committing an offense that presents a threat to public safety, will—consistent with Rule 4.10's decision-making framework—was translated into a risk score. Specifically, Rule 4.10 must mandate that risk assessment reports "explicitly describe the process by which a predicted probability of failure to appear or a predicted probability of rearrest is translated into risk scores."[14] Rule 4.10 must also explain that courts and pretrial risk assessment must not rely on a risk score if not all of the information was available to input into the tool (*e.g.*, if the arrested individual was not able to answer all the questions, as a result of a medical or mental illness), and that a defendant who is unwilling or unable to cooperate by providing information for a risk score must never have that fact used against them.

Rule 4.10 must also mandate that: (a) all system actors receive continual training to ensure consistency, parity, and reliability of risk score calculations, irrespective of race, gender and other immutable characteristics; (b) timely and transparent record-keeping practices that enable the auditing and adjustment of risk assessment classifications over time; and (c) any updates to risk assessment tools are accompanied by a detailed articulation of new intended risk characterizations.[15]

C. Rule 4.10 Must Require Courts to Consider the Factors Considered in the Risk Score, the Weight Given to Each Factor, and How Each Factor Is Defined.

Proposed Rule 4.10 recognizes multiple challenges to the proper use of risk assessment tools stemming from the lack of transparency and explainability of the risk scores. These include the difficulty in confirming the potential inaccuracies of risk scores, the lack of disclosure of information about how some tools weight risk factors as a result of claims from vendors that the information is proprietary, and the need for courts to be familiar with the factors used to determine an individual risk assessment score so as to avoid giving undue weight to those factors when making release and detention decisions.

[14] Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook and Julie Ciccolini, *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, Criminal Justice and Behavior, at 11 (Nov. 23, 2018), available at https://doi.org/10.1177/0093854818811379 (hereinafter "*Layers of Bias*") (collecting studies and research papers).

[15] *See supra*, note 3, Barabas, *et. al*, Open Letter, at 4–5.

The proposal, however, fails to place any limitations on the use of pretrial risk assessment tools to address these concerns. Rule 4.10 must mandate that risk assessment scores be accompanied by individual decision audit trails, which must list all of the factors that went into an individual's the risk score, the weight each factor was given, and how each factor is defined, so that judges can fairly and thoroughly evaluate the risk score in the context of the case at hand and understand what specific factors are being relied upon to define the risk of failing to appear or risk of committing a crime that presents a threat to public safety during the period of pretrial release. Rule 4.10 must mandate that risk assessment scored be accompanied by all of the information necessary to enable a reproducible calculation of a particular individual's risk score.

D. Rule 4.10 Must Ban Judges from Using Proprietary Risk Assessment Tools if a Vendor Refuses to Disclose the Developmental Information About the Risk Score Determination.

Rule 4.10 must require that only pretrial risk assessment tools that provide access to the developmental information about their composition, design, and weighting of factors be used by courts. All of the decisions that went into the development a risk score—*i.e.*, all of the information about the development of the tool and all of the information that goes into a risk score—reflect public policy decisions that will have a direct impact on how the state will proceed in real cases. "Courts need to ensure that researchers, defendants, and judges have access to information that allows them to understand the problems in specific risk-assessment tools, because those problems will be quite different depending on the details of the data, methodology, and conversion to risk scores."[16] Allowing vendors to prevent disclosure of this information through claiming it is "proprietary" also "suppress[es] information to judges themselves, by preventing researchers and defendants from providing fair and thorough evaluations of the risk-assessment instruments."[17] It also keeps defendants from challenging their risk scores and "signal[s] . . . that the government values trade secrets holders as a group more than those directly affected by criminal justice outcomes."[18]

For example, the rate of re-arrest of released defendants could be used as a way to measure someone's risk of committing a crime that threatens public safety if released

[16] *See supra*, note 14, *Layers of Bias*, at 16–17.

[17] *Id*. at 16.

[18] Rebecca Wexler, Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System, 70 Stan. L. Rev. 1343, 1355 (May 2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2920883 (describing the use of trade secret law to shield criminal justice predictive algorithms from legal scrutiny).

prior to trial. But not all jurisdictions define re-arrest in the same way. Some include only re-arrests that actually result in bail revocation, but some include traffic or misdemeanor offenses that don't truly reflect a risk to society. It is important for a court to have a clear understanding of this in order to be able to assess what a risk score actually means. Data about the rate of rearrest is also already "gummed up by our own systemic biases,"[19] and including traffic offenses may make it only more so. Data collected by the Stanford Open Policing Project shows that officers' own implicit or explicit biases cause them to stop black drivers at higher rates than white drivers and to ticket, search, and arrest black and Hispanic drivers during traffic stops more often than whites.[20] Using a rate of arrest that includes traffic offenses could therefore introduce more racial bias into the system, rather than reduce it.

Vendors cannot be allowed to keep such information about the factors that influence risk scores from courts, defendants, and the public. To ensure accountability and the just use of pretrial risk assessment tools, vendors must be transparent about the data and assumptions that they used to build their models and that are reflected in individual risk scores.[21]

E. Rule 4.10 Must Require Courts to Take into Account the Confidence of the Risk Assessment Tool in the Assessed Risk Assessment Score.

Not all risk scores are created equal. The same risk assessment tool may have a high confidence in the risk score that it assigns one defendant, and a low confidence in the risk score it generates for another defendant—even if they are both assigned to the

[19] Sara Chodosh, Courts use algorithms to help determine sentencing, but random people get the same results, Popular Science (Jan. 18, 2018), https://www.popsci.com/recidivism-algorithm-random-bias; *see also* Bernard E. Harcourt, *Risk as a proxy for race: The dangers of risk assessment*, 27 Fed. Sent'g Rep. 237, 238–40 (2015), available at http://fsr.ucpress.edu/content/27/4/237 ("The fact is, risk today has collapsed into prior criminal history, and prior criminal history has become a proxy for race.").

[20] *See* The Stanford Open Policing Project, Findings, https://openpolicing.stanford.edu/findings/; Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Vignesh Ramachandran, Cheryl Phillips, and Sharad Goel, *A large-scale analysis of racial disparities in police stops across the United States* (Working Paper, 2017), https://5harad.com/papers/traffic-stops.pdf.

[21] *See also supra*, note 18, Wexler, 70 Stan. L. Rev. at 1395 (noting that, "when trade secret evidence is relevant to a case, protective orders, sealing, and limited courtroom closures provide sufficient safeguards").

same risk category. Even when placing a simple bet, it would be foolish not to consider the level of confidence in a prediction. When an individual's liberty is at stake, it is improper and irresponsible not to do so. Generalized consideration of the inherent limitations of risk assessment tools is not sufficient. Tables 1 and 2 below provide a simple illustration, by way of example, of how this might occur in practice.

Suppose a tool decides which category to assign to a defendant by first determining the probability that the defendant fits into each of the three categories, "low," "medium," and "high" risk. Then, the tool assigns the defendant to whichever category has the highest probability. In the example illustrated in Tables 1 and 2, two different defendants have both been calculated to be "medium" risk, because for both defendants that category had the highest probability. However, the tool does not have the same confidence for both defendants. In Table 1, the tool has high confidence that the defendant is medium risk, because the probability for "medium" is significantly higher than the probability for "low" or "high." In Table 2, the tool had low confidence that the defendant is medium risk: the probability that the defendant is "medium" risk is only slightly higher than the probability that the defendant is "low" or "high" risk. As a result, it would be improper for a judge to treat both defendants the same based purely on their risk category.

Table 1: High Confidence Prediction of Medium Risk

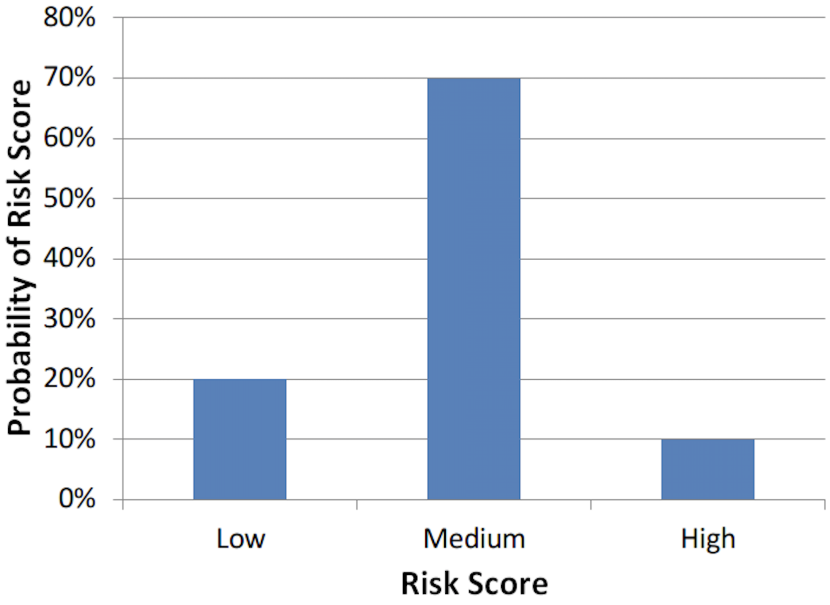
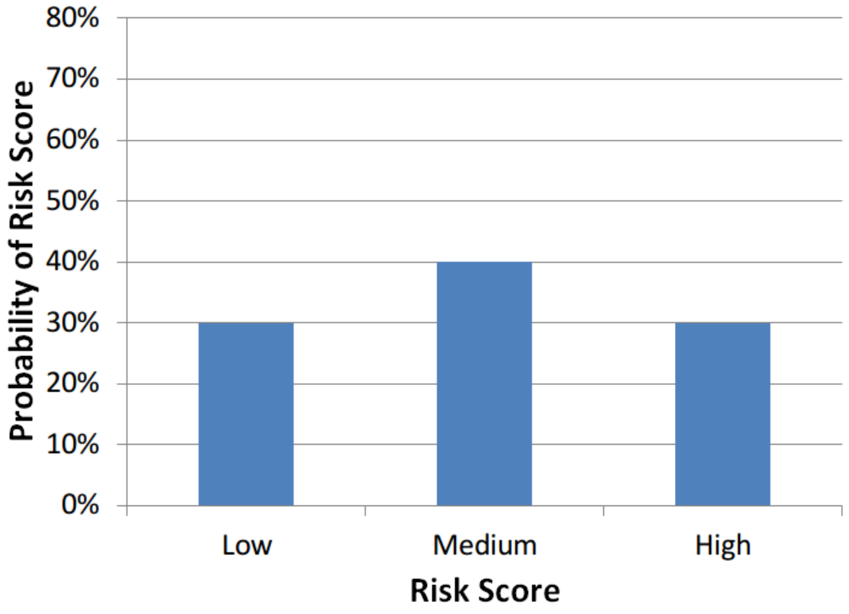


Table 2: Low Confidence Prediction of Medium Risk



Rule 4.10 must mandate that, particularly when a defendant be assessed medium or high risk, courts take into account the validated statistical confidence of the individual risk assessment score given to the defendant in the case at hand, as well as the specific factors and/or rationale informing the tool's confidence in the risk assessment.[22] In order to ensure that courts can properly assess the statistical confidence reported, Rule 4.10 must also mandate that risk assessment reports provide a graphical representation of the risk assessment data that illustrates any limitations inherent in the prediction.

F. Rule 4.10 Must Define Processes by Which Defendants Shall Get Access to the Risk Assessment Report—Including their Risk Scores, All of the Information that Went into Their Scores, and All Information Regarding the Tool's Confidence of the Risk Scores.

To ensure that defendants can meaningful evaluate and challenge their risk scores, 4.10 must define the specific processes by which defendants shall be provided access to their risk assessment reports. To meaningfully challenge their risk scores, defendants must have access to all of the information that a judge has access to, pursuant to our recommendations—*i.e.*, the entire risk assessment report and any and all information

[22] *See supra*, note 14, *Layers of Bias*, at 11.

related to the results of the risk assessment, including all of the information that went into defining and calculating their risk scores and any information regarding the tool's confidence in their risk scores.

Rule 4.10 must also mandate that defendants shall be granted access to the training data, validation data, error rates, and the results of any and all fairness and bias evaluations, so that they may meaningfully evaluate and challenge their risk scores. When necessary to protect the privacy of other defendants, Rule 4.10 should require that data unrelated to, but required to be released to a particular defendant be anonymized and/or subject to a protective order.

G. Rule 4.10 Must Define a Procedure for How Defendants Can Challenge the Risk Assessment Tool and their Resulting Score.

To ensure that defendants' due process rights are protected, Rule 4.10 must also mandate that defendants be provided a meaningful opportunity to challenge both the risk assessment tool and its resulting risk score. Rule 4.10 must lay out the specifically procedures by which a defendant may do so. Ensuring that risk assessment tools and individuals risk scores can be assessed via an adversarial process is a critical safeguard for identifying and rectifying mistakes that will inevitably occur—whether they are caused by biased data, biased models, or inaccurate input variables—as well as potential manipulation of risk scores by court officers and staff.

H. Rule 4.10 Must Include the Language Considered by the Judicial Council Limiting the Use of the Pretrial Risk Assessment Information in Any Further Proceedings.

The Judicial Council considered, but decided against, including in Rule 4.10 a requirement that a risk score not be used "for any purpose other than a determination of pretrial release or release or detention in the current proceeding, or conditions of release, unless both parties otherwise stipulate."

The Judicial Council must include this limitation in Rule 4.10. The defendants' risk score for failing to appear for their court date and separate risk score for committing a crime that presents a threat to public safety during pre-trial release must be based solely on the risk that the defendant will not show up or commit such a crime during the specific period at issue. The formulas underlying these tools must be tailored to the specific request at hand, and the scores generated must be used only for that specific purpose. Indeed, predicting these specific risks in the pretrial setting is distinct from predicting them in the context of parole, probation, or sentencing, and a tool must be validated for predicting risk in each specific context. The risk score generated in one context should therefore not be reused in another context. Rule 4.10 must therefore prohibit the use of the risk assessment information for any other purpose, unless consented to by the defendant.

I. Rule 4.10 Must Require that Judges Consider Studies/Reports from Independent Groups About a Tool's Bias and Fairness.

Rule 4.10 provides that the use of pretrial risk assessment information is meant to "[a]ddress any biases in pretrial release and detention decision-making." Proposed Rule 4.10(a)(2)(C). SB 10 itself mandates that any risk assessment tool be "demonstrated by scientific research to . . . minimize bias." Cal. Pen. Code § 1320.7(k).

Yet, the proposed rule fails to place categorical limitations on the use of pretrial risk assessment tools if these criteria are not satisfied, or mandate that judges consider data documenting a specific tool's potential biases when making a decision based on a risk score. It is not enough for courts to merely consider as factors whether or not any scientific research has raised questions about an instrument's fairness or bias, or whether or not a particular instrument has been validated at the local level, as the proposed rule currently contemplates. *See* Proposed Rule 4.10(b)(5)(C), (D). Studies have demonstrated that these tools—as a result of the very biases baked into the criminal justice system that they are intended to correct—disproportionately impact communities of color and reinforce existing inequalities. A pretrial risk assessment tool not magically solve deeply-rooted bias and inequity in the criminal justice system—particularly not a tool built on biased data. Consistent with both the purpose statement articulated in Rule 4.10 and the text of SB 10 itself, independent scientific research regarding discriminatory and biased outcomes, as well validation of the tool at the local level, must be mandated prior to the implementation of the pretrial risk assessment tool by any court representative.

While the rule of court promulgated pursuant to Cal Pen. Code § 1320.24(a)(2) will address the specific procedures by which tools should be validated and by which bias should be identified and mitigated, Rule 4.10 must explicitly mandate that courts not use information generated from a pretrial risk assessment tool unless and until it has been evaluated and found to satisfy fairness criteria demonstrated by independent scientific research to mitigate bias—and that any results of such review be published and shared with defendants. In fact, a recent research paper distills three distinct forms of bias inherent in pretrial risk assessment tools that have been identified in the academic literature: (i) bias embedded in the statistical model underlying the tool; (ii) bias embedded in the data used to create the risk assessment model; and (iii) bias implicit in data-driven risk assessment scoring, where decisions about individuals are based on groups.[23] It is improper and irresponsible for courts to rely on pretrial risk assessment information from tools that have not been transparently evaluated on each of the three bias layers—both initially and on a continuing basis.[24]

Specifically, Rule 4.10 must require that judges consider (and that defendants, and independent researchers, possibly operating under the supervision of an institutional review board, be granted access to):

(a) the methodology and results of an evaluation by an independent group (*i.e*., not the vendor) on each layer of bias, including an assessment of the false positive and false negative rate across groups;

(b) the results of an assessment by an independent group about whether the tool satisfies fairness criteria, selected by the Judicial Council with the opportunity for public comment pursuant to Cal Pen. Code § 1320.24(a)(2), that have been demonstrated by independent scientific research to mitigate bias on the basis of race, gender, or other protected classes;[25]

(c) the results of a validation check conducted on the tool within the last 12 months "by an independent group—possibly a standing commission—which includes perspectives of statisticians, criminologists, and pretrial and probation service workers specific to the relevant jurisdiction";[26] and

(d) the results of tests mandated on a regular basis, as outlined by the Judicial Council with the opportunity for public comment pursuant to Cal Pen. Code § 1320.24(a)(2), measuring the disparate impact of tool error rates by race, gender, and other protected classes.

There are various ways of measuring whether a model makes fair predictions, and those methods are fundamentally incompatible: "Resolving bias in one way produces a different type of bias."[27] The decision about what fairness metric to use to evaluate any given tool—and how to weigh the competing trade-offs between the various metrics—is a public policy decision that must be made by the Judicial Council, with the opportunity

[23] *See supra*, note 14, *Layers of Bias*.

[24] Research shows that risk assessment tools must be evaluated regularly and repeatedly over time to ensure their validity. *See supra*, note 9, Risk and Needs Assessment and Race.

[25] *See supra*, note 3, Barabas, et. al, Open Letter, at 3.

[26] *Id*.

[27] *See supra,* note 14, *Layers of Bias,* at 5.

to public comment, pursuant to Cal Pen. Code § 1320.24(a)(2). But because the fairness metric will have an impact on risk assignments assigned to individuals, "[d]efendants and defense lawyers" and courts "should be able to analyze model fairness—and the criteria used to measure fairness—to make liberty determinations about their cases."[28]

Indeed, some fairness metrics might actually reinforce bias—rather than reduce it—when applied in the context of pretrial risk assessment, and courts and defendants must have access to that information in individual cases to be able to properly assess the information produced by the tool. For example, studies suggest that the predictive parity model, which assesses whether the predictive value of a risk score is similar across groups, will only ensure fair results if "the criminal justice system is equally fair for White and Black people[.]"[29] We know, however, that this is unlikely to be true. Studies document that communities of color are more likely to be cited, arrested, prosecuted, and wrongfully convicted by police.[30] Models based on existing data will therefore "make people of color look riskier than Whites" and "the predictions are necessarily biased."[31] And one of the key rationales for adoption of risk assessment tools is the need to counteract bias and inequality in the system.

The problem with comparing predictive values in the context of pretrial risk assessment lies in the fact that models assume that the future will be like the past. Criminal risk assessment models "are trained on data generated by past police bias" and we ask the models "to predict events that are dependent on future police bias"; "[w]hen both the data used to produce the risk-assessment instrument and the data used to evaluate it come from the criminal justice system, quantitative risk assessments merely *launder that bias*."[32]

Other fairness metrics, however, have been demonstrated by independent scientific research to minimize bias, as SB 10 requires. One such metric considers false positive rates (whether defendants from one group who do not reoffend are more likely to get high risk scores than defendants from other groups who likewise do not reoffend), false negative rates (whether people who do go on to commit crimes get similar scores

[28] *Id.* at 11.

[29] *Id*. at 20.

[30] *See* The Sentencing Project, Report of the Sentencing Project to the United Nations Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia, and Related Intolerance Regarding Racial Disparities in the United States Criminal Justice System (Mar. 2018), available at https://www.sentencingproject.org/publications/un-report-on-racial-disparities/.

[31] *See supra,* note 14, *Layers of Bias,* at 13.

[32] *Id.* at 8 (emphasis added).

across groups), or the error rate balance (assessing whether false positive and false negative rates are equal across groups).[33] Studies show that differences in false positive and false negative rates across groups "can result in disparate impact under policies where a high-risk assessment results in a stricter penalty for the defendant"—including when the results are used to inform bail decisions.[34] Thus, risk assessment information must not be used in courts until *at a minimum* the false positive and false negative rates across groups have been evaluated and that data made available to the court, individual defendants, and the public.[35]

II. Rule 4.40: Review and Release Standards for Pretrial Assessment Services, for Persons Who Have Been Assessed as Medium Risk

Rule 4.40 is meant to prescribe the parameters of local rules adopted pursuant to SB 10 setting forth pre-arraignment "review and release standards" for those assessed as medium risk[36] and eligible under SB 10 to be released on their own recognizance or on supervised own recognizance. *See* Cal. Pen. Code §§ 1320.11(a), 1320.24(a)(4). The proposed rule, however, provides no meaningful parameters on the use of pretrial risk assessment information, and fails to place any meaningful limitations on the discretion of Pretrial Assessment Services to detain individual defendants.

A. Rule 4.40 Must Mandate that Pretrial Assessment Services Only Use Risk Assessment Information to Recommend Release, or to Aid in the Assessment of What Conditions of Release are Appropriate.

While SB 10 enables Pretrial Risk Assessment Services to consider whether to detain a people who are assessed medium risk, the decision to detain must be based entirely on other factors—not the risk assessment score/information.

[33] *See supra,* note 14, *Layers of Bias* at 5–6.

[34] *See also* Alexandra Chouldechova, *Fair prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, at 4 (last revised Feb. 28, 2017), https://www.andrew.cmu.edu/user/achoulde/files/disparate_impact.pdf.

[35] Note that when recidivism prevalence differs across groups—as it almost always, if not always, does due to inherent biases in the criminal justice system and discrepancy in access to resources—in fairness measures based on predictive parity are incompatible with measures based on equal false positive and negative rates across groups. *Id.* at 5–7.

[36] Consistent with our comment in response to Rule 4.10, Rule 4.40 must only apply if the defendant's medium risk assessment is consistent with the thresholds for low, medium, and high risk that the Judicial Council sets pursuant to specific policy objectives of the State.

First, because of the inherent limitations and challenges of risk assessment tools—including that they are designed to predict a likelihood of the risk of groups, not individuals, and cannot predict the future behavior of any particular individual[37]—they are not appropriate for recommending detention or for informing decisions about whether to detain an individual.

In addition, allowing Pretrial Assessment Services to rely on the risk score to inform a decision to detain an individual will guarantee that Pretrial Assessment Services gives additional undue weight to the factors identified in Rule 4.40(b)(3), as each of these factors will likely be represented in the individual's risk score. Rule 4.40 must mandate that a decision to detain an individual who is scored as "medium" risk be *independent* of the risk score.

Rule 4.40 must therefore mandate that Pretrial Assessment Service use risk assessment information *only* for the purpose of recommending release and informing appropriate conditions of release. Rule 4.10 must also specify that where the validated statistical confidence of a medium risk score shows that the model is only marginally confident that the defendant falls within the medium risk category, as compared to the low risk category, it should weigh in favor of release.

B. "Criminal History," as Used in Rule 4.40(b)(3)(B), Must Exclude Arrests that Did Not Result in Convictions.

For the same reason that pretrial risk assessment tools must not consider arrests that did not result in conviction in assessing the risk that a defendant is likely to commit a violent offense in the pretrial period that justifies pretrial detention, Pretrial Assessment Services must not consider arrests that did not result in conviction in deciding whether to detain pursuant to Cal. Pen. Code § 1320.10(c).

Pretrial Assessment Services must make an *independent assessment*, irrespective of the risk score, of how likely a defendant is to commit a violent offense in the pretrial period that justifies pretrial detention—not whether the defendant is likely to be convicted of or arrested for any other alleged offense in the future. As noted above, rearrest is not a good proxy for future criminality; it is well documented that different demographic groups are stopped, searched, arrested, and charged at very different rates across the United States.[38] Thus, "criminal history," as used in Rule 4.40(b)(3)(B), must exclude arrests that did not result in convictions.

[37] *See* Proposed Rule 4.10(b)(5)(A).

[38] *See supra*, note 6.

C. The Judicial Council Must Require that the Reports Required Pursuant to Proposed Rule 4.40(e)(2) Be Released to the Public and Available Via Each Court's Website.

To ensure transparency and accountability, and to enable independent research and review of the use of risk assessment tools across the state, Rule 4.40 must be updated to mandate that all reports submitted to be Judicial Council pursuant to subsection (e)(2) be published publicly.

D. Rule 4.40 Must Mandate a Process for Identifying Exclusions that Involves Consultation with Stakeholders

Rule 4.40 must also be updated to provide specific procedures to ensure that community members and system stakeholders have a role in decision-making on these exclusions. These procedures must guarantee the opportunity for robust public comment and ensure transparency and accountability.

***

Thank you for this opportunity to submit comments on Proposed Rules 4.10 and 4.40. If you have any questions, please feel free to contact us at either jamie@eff.org or +1 (415) 436-9333, x164.

Respectfully Submitted,

Jamie Williams, Staff Attorney
Stephanie Lacambra, Criminal Defense Staff Attorney
Jeremy Gillula, Tech Projects Director
Electronic Frontier Foundation
815 Eddy Street
San Francisco, CA 94109
United States of America
Telephone: +1 (415) 436-9333
Email: jamie@eff.org