Primer on Risk Assessment Instruments for Legal Decision-Makers

CHRISTOPHER SLOBOGIN
Milton R. Underwood Chair in Law

VANDERBILT®
LAW SCHOOL

Funded by

Charles Koch
CHARLES KOCH FOUNDATION

A PRIMER ON RISK ASSESSMENT INSTRUMENTS FOR LEGAL DECISION-MAKERS

Preface

This primer is addressed to judges, parole board members, and other legal decision-makers who use or are considering using the results of risk assessment instruments (RAIs) in making determinations about post-conviction dispositions, as well as to legislators and executive officials responsible for authorizing such use.[*] It is meant to help these decision-makers determine whether a particular RAI is an appropriate basis for legal determinations and whether evaluators who rely on an RAI have done so properly. This primer does not take a position on whether RAIs should be integrated into the criminal process. Rather, it provides legal decision-makers with information about how RAIs are constructed and the types of information they provide, with the goal of facilitating their intelligent selection and use.

The work on the Primer was funded by the Charles H. Koch Foundation. It involved consultation with a number of experts on risk assessment, including Dr. Sarah Desmarais, a psychologist and professor at North Carolina State University; Brandon Garrett, a law professor at Duke Law School; Melissa Hamilton, a law professor at the University of Surrey; Dr. Rhys Hester, a sociologist, lawyer and professor at Clemson University; Cecelia Klingele, a law professor at the University of Wisconsin; Sandra Mayson, a law professor at the University of Georgia Law School; Dr. John Monahan, a psychologist and professor at the University of Virginia Law School; Michael O'Hear, a law professor at Marquette Law School; Kevin Reitz, a law professor at the University of Minnesota Law School; Dr. Jennifer Skeem, a psychologist and professor at the University of California, Berkeley; and Dr. Megan Stevenson, an economist and professor at the University of Virginia Law School. However, the ultimate responsibility for this work falls on the author of the Primer, Christopher Slobogin, law professor at Vanderbilt University. Any errors in concept or fact are his, and his alone.

[*] Much of what is said here could be relevant to use of risk assessment instruments during the pretrial process as well, but this primer is directly solely at decision-makers involved with risk assessment in the post-conviction process.

TABLE OF CONTENTS

WHAT ARE RISK ASSESSMENT INSTRUMENTS?

The term "risk assessment" encompasses a number of different practices that differ in the extent to which they: (1) rely on empirically-valid *risk* and *protective* factors (i.e., factors that statistically correlate with an elevated or reduced risk of reoffending); (2) use a structured method for measuring these risk and protective factors; (3) establish a procedure for combining scores on the individual risk and protective factors into a total score; and (4) produce a final estimate of risk.[1] *Clinical* risk assessment—in many settings, the traditional and still typical method used by many judges, parole boards and mental health professionals—structures none of these components; rather an estimate of risk is based on experience and perhaps intuition, and the factors considered may vary from case-to-case, and be applied differently in different cases. *Checklist* risk assessment provides structure on the first component, by listing the factors that should be considered. *Structured professional judgment* (SPJ) risk assessment satisfies the first two components, by providing a list of factors and indicating how they should be measured (e.g., on a scale of 0-2), but avoids combining these measures for a total score, instead counselling that the item ratings be considered merely in arriving at an overall conclusion about risk. *Adjusted actuarial* risk assessment lists the factors, describes how they should be measured, and produces a total score, but allows evaluators to adjust or modify the score based on clinical judgment that is not structured by the instrument. Stand-alone *actuarial* risk assessment does not permit such adjustments, but rather produces a probability estimate that is considered final.

The term "risk assessment instrument," as used here, applies only to the last three types of practices. An example of an SPJ instrument is the HCR-20, version3 (hereafter, HCR-20), which is composed of 20 items, rated on a scale of 0-2.[2] Examples of adjusted actuarial instruments are the Correctional Offender Management Profiling for Alternative Sanctions tool (COMPAS),[3] and the Legal Service/Case Management Inventory (hereafter, LSCMI).[4] Examples of actuarial instruments are the Violence Risk Appraisal Guide-Revised (hereafter, VRAG),[5] Virginia's Non-Violent Risk Assessment (NVRA),[6] and the STATIC-99.[7] These examples are fluid, however. For instance, if evaluators so choose, the LSCMI can be used as an actuarial instrument, and the scores on the VRAG and STATIC-99 can be adjusted.

RAIs are based on research that tries to identify factors that correlate with risk. Typical risk factors in RAIs include criminal history, diagnosis, criminal attitudes, school and work status, and substance use.[8] Risk factors are sometimes distinguished in terms of whether they are "static" (unchanging) or "dynamic" (changeable). Some RAIs also claim to identify "needs," which, properly defined, are dynamic factors that, if changed, reduce risk. For instance, employment status and substance abuse are dynamic factors that can be helpful in estimating risk but, to date, only substance abuse has clearly been shown to be a criminogenic need—that is, one that, if successfully treated, reduces risk.[9]

CONTROVERSIES OVER RISK ASSESSMENT INSTRUMENTS

As indicated in the preface, this primer does not take a position on whether RAIs should be used in the criminal justice system. However, legal decision-makers who use or are considering using these tools should have some understanding of the criticisms of RAIs and how RAI proponents have responded to these criticisms. The criticisms, and the responses to them, are organized here under six headings: assessing individual cases based on group data; the helpfulness of RAIs to lay decision-makers;

racial disparities; use of uncontrollable factors; dispositional consequences; and procedural concerns.

Assessing Individual Cases Based on Group Data

RAIs—especially those that are actuarial—have been criticized on the ground that research about groups cannot predict whether a given individual will reoffend. As one judge put this point, "[n]ot only are . . . statistics concerning the violent behavior of others irrelevant, but it seems to me wrong to confine any person on the basis not of that persons' own prior conduct but on the basis of statistical evidence regarding the behavior of other people."[10] This issue has been called the "G2i" problem, the application of general information to individual cases.[11]

The statistical argument purporting to support the stance that G2i is not possible has been debunked by noted statisticians Peter Imrey and Philip Dawid, who find it "seriously mistaken in many particulars" and who conclude that it should "play no role in reasoned discussions about violence recidivism risk assessment."[12] Other commentators have pointed out that, if these assertions were true, any type of expert testimony—whether or not empirically based—would be suspect.[13] Even a clinical risk assessment by a judge or mental health professional relies—consciously or not—on stereotypes, past experiences with "similar" individuals, and lessons learned from the literature about groups. As one commentator noted: "Although the clinician need not identify in advance the characteristics he will regard as salient, he must nevertheless evaluate the applicant on the basis of a finite number of salient characteristics, and thus, like the statistical decisionmaker, he treats the applicant as a member of a class defined by those characteristics."[14]

However, the fact that the G2i phenomenon is unavoidable does have implications for how to frame prediction results. A conclusion based on an RAI (or any other conclusion about risk) should not purport to say that a person will, or will not, reoffend—a fact that, in almost all cases, is unknowable. Nor should it state that a particular offender has X probability of reoffending. Rather, an evaluator using an RAI to estimate the risk an offender poses should report that the offender received a risk score that is consistent with the scores of a group, X percent of which offended in the past.[15]

The Helpfulness of RAIs

The statistical expertise needed to put together an actuarial RAI and the familiarity with research literature that goes into the creation of SPJ instruments requires a significant amount of specialized knowledge. But if judges or lay-parole boards are able to figure out who will reoffend as accurately as evaluators using RAIs, then arguably that specialized knowledge is not helpful to the factfinder. One study purported to find that the COMPAS "is no more accurate or fair than predictions made by people with little or no criminal justice expertise."[16] Specifically it found, based on a comparison of COMPAS and human predictions for 1,000 defendants, that while the COMPAS was correct in 65% of the cases, humans were correct in about 62% of the cases.[17]

However, others have pointed out that the 50 mini-vignettes shown to the lay subjects in this study listed only a few features of the defendant, all of which have a robust statistical relationship with reoffending; in effect the subjects were provided a "checklist" RAI. Further, immediately after each prediction, the subjects were told whether they were right or wrong, feedback that a judge or parole board never receives but which, in the study, "trained" the participants about the most pertinent traits and how they are related to recidivism.[18] A

follow-up study found that when lay participants are *not* provided feedback, they do much more poorly than an RAI, even when they are given base rate information about the average rate of offending of the population in question.[19] The authors of this study also found that when the information given the humans was "noisier" (that is, when a much richer set of facts was provided than the barebones list of traits provided in the original study), the lay subjects did barely better than chance, whereas the statistical model the authors created was much better at distinguishing recidivists and non-recidivists.[20]

These types of results replicate a large number of other studies finding that algorithms typically outperform human predictions, whether the latter are made by laypeople or trained clinicians.[21] For instance, a 2006 meta-analysis of 41 studies found that actuarial techniques routinely did better than clinical methods in every area investigated, and that with respect to predicting violent or criminal behavior in particular, the actuarial approach was "clearly superior to the clinical approach."[22] The study also found that even subsets of "best professionals" designated as experts did not outperform statistical formulae.[23] Several studies that compare algorithms to judges, clinicians, and correctional officers obtain similar results, probably because, despite their official position, the decisions of these groups are often like other peoples'—largely intuitive, heuristic-based, subject to bias, and inattentive to base rates.[24] This research calls into question both clinical risk assessment and actuarial assessment that is adjusted, at least when the adjustments are not based on empirical research or on obvious anomalies (such as when a person designated as high risk for violence has since become disabled, or when a person considered low risk voices a genuine threat to another).[25]

Of course, the advantages of RAIs over clinical judgment are lost if they are poorly constructed. The validity of an RAI can be measured in several ways. Further, RAIs must also be validated and re-validated on the relevant populations. These means of gauging accuracy are described further below.

Racial Disparities

Risk assessment instruments might be considered unfair on a number of grounds. Most prominently, critics have argued that RAIs can perpetuate racially disparate effects. For instance, RAIs may be developed using criminal arrest history that, because of racialized police practices, overestimates the actual crime rates by people of color, and therefore may produce results indicating that this group is higher risk than it actually is.[26] Further, research has shown that even RAIs that assign risk levels relatively accurately across races can produce higher "false positive rates"[†] for blacks than for whites, and higher "false negative rates"[‡] for whites than blacks; if so, a higher proportion of blacks could be wrongly incarcerated or receive enhanced sentences, while a higher proportion of whites could be released and reoffend. More generally, risk factors such as employment status, socio-economic status, or location may correlate with race, which some critics argue should render reliance on such factors impermissible, and perhaps even unconstitutional.[27]

To some extent, the effects of racialized policing can be neutralized. For instance, some algorithms refrain from predicting arrests for minor drug offenses and non-violent mis-demeanors, while others develop separate risk categories for different racial groups, and still

[†] The percentage of people who did not recidivate but had been predicted to recidivate (i.e., given a positive prediction).

[‡] The percentage of people who recidivated but had been predicted to not recidivate (i.e., given a negative prediction).

others exclude risk factors that are not equally predictive across racial groups.[28] Unfortunately, higher false positive rates for blacks are inevitable as a statistical matter if, as some research shows,[29] the recidivism rate is higher among blacks than whites.[30] Further, adjusting scores in an effort to equalize false positive and false negative rates across races will decrease the accuracy of RAIs by over-estimating the risk posed by whites and under-estimating the risk posed by blacks; in any event, simultaneous equalization of both false positive and false negative rates may be impossible.[31] And under current constitutional doctrine, unintentional disparate racial impact does not violate the equal protection clause,[32] especially if the risk factors that correlate with race meaningfully improve accuracy and thus help achieve the state's interest in protecting the public.[33]

Finally, consistent with the research showing that RAIs are generally more accurate than clinical judgment, several studies indicate that human adjustments to RAIs increase racially disparate outcomes.[34] It may be that structured risk assessment is better at avoiding such outcomes than unaided human decision-makers.[35] Further, because the workings of an algorithm are more transparent than those of the human mind, if racially disparate impacts are discovered, they are more easily corrected using RAIs.[36] As Sendhil Mullainathan has noted, "biased algorithms are easier to fix than biased people."[37]

Use of Uncontrollable Factors.

A related criticism of RAIs is that many of their risk factors do not describe blameworthy conduct. For instance, although no RAI explicitly uses race as a risk factor, several use sex (with maleness increasing the risk score), age (with youth increasing the risk score), diagnosis (with certain personality disorders increasing the risk score), and early childhood experiences (with abuse and parental absence increasing the risk score). Critics argue that punishment should not be enhanced because of such factors, or even by factors such as "criminal attitudes" and impulsivity, since they are not per se criminal.[38]

At the same time, the courts have long recognized that post-conviction dispositional decisions—including the death penalty—may be based in part on a risk assessment;[39] if so, a risk-based disposition that is not prolonged beyond what an offender fairly deserves under the relevant sentencing scheme should be permissible even if based in whole or part on non-blameworthy factors.[40] It has also been argued that risk-based dispositions are imposed not *because* a person is young, male, or has a particular personality disorder, but because these factors, along with others, indicate that the person is more (or less) likely to choose to commit a blameworthy act in the future.[41] Viewed this way, risk factors are merely circumstantial evidence about future blameworthiness. Finally—repeating the comparative theme—even non-empirical risk assessments undoubtedly rely, consciously or not, on such factors.

It is sometimes stated that, as long as a risk assessment is used only to identify lower risk offenders who can be given a reduced sentence or diverted from imprisonment altogether, concerns about the use of particular risk factors are alleviated.[42] However, the tension between risk and desert is not that easily resolved. For instance, a jurisdiction that lowers the sentence of an older woman who is married based on those factors is, in effect, raising the sentence of a younger, single male. A regime that favors low risk offenders automatically disfavors high-risk offenders.

Dispositional Consequences

An over-arching concern about using RAIs is that they will make risk a more salient consideration during the post-conviction

process, to the detriment of achieving other punishment goals.[43] The focus will become reduction of recidivism rather than proportionate punishment based on just desert, general deterrence, rehabilitation or other, more broad-based efforts to reduce crime through social programs. Further, that focus could lead to an increase in imprisonment and sentence lengths, given the fact that all offenders pose some risk if released. Finally, there is the concern that, because they control the construction of RAIs, tool developers, not policymakers, will control the definition of risk.[44]

Proponents of RAIs respond that, while all of this is a possible effect of increased reliance on RAIs, the more likely impact is a reduction in incarceration rates and sentence lengths, and an increase in rehabilitative services.[45] For instance, researchers with bipartisan credentials who audited the compositions of the prison populations in three states estimated that, if danger to the community were the only justification for continued confinement, roughly half the prisoners would be released.[46] If low risk is defined as those posing a risk of violent crime below the average risk posed by offenders, and this group is either not imprisoned or imprisoned for the minimum term, incarceration populations are likely to be significantly reduced, for people of color as well as other groups;[47] research shows that recidivism among low risk individuals actually diminishes if imprisonment and enhanced surveillance are avoided.[48] Further, to the extent RAIs identify criminogenic needs, they can facilitate assignment to treatment programs. As these examples illustrate, policymakers, not RAI developers, must dictate the specifications of the instruments.[49]

Procedural Concerns

Even if RAIs are relevant and helpful, and their use does not violate equality or punishment norms, some critics express worry that the quantified nature of RAIs, and the fact that their inner workings are not easily understood, undermines the procedural fairness of the risk determination.[50] Basing a risk assessment on a finite number of factors and, when the assessment is actuarial in nature, on a risk or probability score, may seem antithetical to an "individualized" evaluation of the person. Further, to the extent the algorithm is not transparent—which may occur either because it was developed by a private company claiming trade secret protection or because the RAI is based on deep machine learning (artificial intelligence) techniques—independent due process concerns arise.[51] These challenges are exacerbated if, as is true in most jurisdictions, the rules governing admissibility of scientific evidence, such as those developed under *Daubert v. Merrell Dow Pharmaceuticals*,[52] do not apply post-conviction, and if, as is true with parole hearings, the proceeding is informal, with no right to counsel.

The extent to which these challenges can be overcome depends upon the law and practice in a given jurisdiction. At a minimum, the subject of the risk assessment should always be able to contest the accuracy of conclusions that a given risk factor (such as a particular arrest) applies, as well as the way in which the risk factors are combined. Legal decision-makers can assist in this process, especially when counsel is not present. Explanations of the risk assessment process should also be provided, and the workings of the RAI should be made as transparent as possible. Perhaps most importantly, the validity of any given RAI used by the state should be subject to peer review and testing before it is relied on to make legal decisions, which can help minimize concerns about the informality of the post-conviction process. These matters are discussed further below.

GENERAL PRINCIPLES OF RISK ASSESSMENT

With these controversies in mind, and as explained further in the next section, several general principles should govern the criminal justice system's reliance on risk assessment instruments (RAIs) during the post-conviction process.

1. Legal policymakers (legislatures, executive agencies or courts) should define the threshold and nature of the risk that is relevant for each risk assessment setting (e.g., sentencing, parole release, within-in prison management). Policymakers should not cede this authority to the developers of RAIs.
2. Developers of RAIs should provide validation data indicating: (a) the proportion of people within each risk category or with a given risk score who reoffend, as defined under principle 1; (b) the ability of the RAI to discriminate between recidivists and non-recidivists above chance levels; and (c) inter-rater reliability (agreement between different evaluators).
3. The validation data should, if possible, be based on a population from the jurisdiction in question and report data about the validity measures described in principle 2 for men, women, different ethnicities and other important demographic groups.
4. Jurisdictions should verify the predictive validity and inter-rater reliability data described in principles 2 and 3 through a peer-review process that is independent from the developer of the RAI and that has access to the developer's inputs, statistical analysis, and outcome data.
5. Jurisdictions should re-validate the RAI after significant changes in the relevant population, the rate of crime, and norms or policies that are likely to affect recidivism rates substantially.
6. Risk estimates based on an RAI should be expressed in terms of membership in a group with a specified recidivism rate, not in terms of whether a particular person will or will not recidivate.
7. In estimating risk, the results of a validated risk assessment instrument are preferable to a clinical judgment about risk and should be given presumptive, but rebuttable, effect.
8. Legal decision-makers and evaluators should be trained in the use of RAIs to the extent they are responsible for obtaining and analyzing their results.
9. In individual cases, the risk factors and the way in which they influence the ultimate conclusion on risk should be available to the parties and subject to adversarial testing.
10. The results of a risk assessment should not lengthen a sentence beyond the maximum punishment that the legislature has determined is appropriate for the offense of conviction.
11. To the extent consistent with just desert, general deterrence and other punishment considerations, decision-makers charged with responding to a risk assessment should impose the least burdensome measures that can sufficiently address risk, which may mean, for instance, that offenders considered to be lower risk are subject to little or no intervention.
12. When consistent with other goals of punishment, legal decision-makers should consider supportive measures to reduce risk, including access to treatment, education or social services. To this end, RAI developers should endeavor to include variable or dynamic risk factors that have been shown empirically to predict re-offending and to be changeable through intervention.

The queries highlighted below elaborate on these principles.

ISSUES THAT LEGAL DECISION-MAKERS USING RAIS SHOULD CONSIDER

The following material expands on these principles, organized around 15 questions about RAIs that relate to their (1) relevance, (2) accuracy, and (3) fairness. Whether these queries are resolved on a case-by-case or jurisdiction-wide basis will depend upon the law and practice of the jurisdiction.

Relevance/Fit

In evaluating the results of an RAI, the legal decision-maker should ascertain whether the information it provides is pertinent to the legal issue in question, with respect to: (1) the probability of reoffending; (2) the type of reoffending predicted; (3) the time frame within which the reoffending will occur; and (4) the intervention(s) that are likely to reduce the predicted risk.

1. *Does the RAI provide information about the **probability** that a person within the individual's risk category will recidivate?*

As noted above, the most accurate way to designate a person's risk level is by associating the person with a group, as in "research indicates that approximately 30% of the people with this person's risk score recidivated in the absence of risk-reducing intervention." This probability is also the most useful information for the binary decision that a judge makes about whether a person should be sent to prison or diverted to an alternative, or that a parole board makes in deciding whether to release a prisoner. Ideally, as principle #1 indicates, the legislature or appellate jurisprudence would assist in that endeavor by providing probability thresholds appropriate for the legal setting (which might vary depending on whether sentencing, parole, or within-prison dispositions are involved), after considering the associated false positive and false negative rates, and the accompanying imprisonment and treatment costs.

Unfortunately, in-depth analysis of this sort has, to date, been rare. To the extent a jurisdiction's law defines dangerousness in the sentencing context, it is usually very imprecise. For instance, the relevant law might describe the relevant threshold as a "probability" or "significant possibility" of reoffending, without further elaboration.[53]

With the advent of RAIs, risk can be quantified more precisely. Not all RAI developers take advantage of this capability, however. Instead, evidence of risk is often limited to statements about whether the offender is "high risk," "medium risk" or "low risk" (as occurs under the HCR-20) or a description of the offender's "risk decile" (as provided by the COMPAS). These statements do not reveal the specific group-based probabilities associated with, respectively, the instruments' risk designation or decile.

Legal decision-makers should seek out such information rather than rely on categorical designations such as "high" or "low" risk. Otherwise, there is a significant chance they will be misled. For instance, surveys that ask clinicians and judges how they define "high risk" show tremendous variation. The average percentage associated with that phrase falls somewhere between 60 and 70%, but the range of answers varies from 5% to 100% and the variability between raters is very high.[54] One study of evaluators found that the probability of recidivating associated with a "moderate-high" rating was more than twice the actual recidivism rate of those groups.[55] Similarly, the decile designation does not provide sufficient information about probabilities. For instance, "third decile" on the COMPAS does not refer to a group, 30% of whom will recidivate, but rather a group that, in the validation sample, posed a lower risk than roughly 70% of the sample; because even the highest decile group on the COMPAS is associated with a probability of

reoffending considerably lower than 100%, the third decile group is very likely associated with a reoffending risk much lower than 30%.

2. *Is the **type of conduct** that the RAI predicts relevant to the legal decision at issue?*

Just as the law often only vaguely defines the probability threshold, it may not indicate the outcome measure or outcome variable of interest for the legal setting in question. Should the person be considered "dangerous" for legal purposes if it is shown, with the requisite probability, that someone in the person's risk category will be *arrested* for *any* offense? Or only if the person will be *convicted* for a *violent* offense? And, if the latter, how is violence defined? These questions are not always answered in the relevant legislation or caselaw. If not, legal decision-makers evaluating individual cases must make their own judgment about this issue.

While many RAIs provide probabilities for both "general recidivism" and "violent recidivism," many define those terms very broadly. General recidivism, for instance, could be defined in terms of *arrest* for *any* crime (as is the case for the COMPAS, for instance). Other RAIs might include as the outcome variable infractions of prison disciplinary rules. In such cases, the legal decision-maker must ask whether the probability of re-arrest for a misdemeanor, or the commission of any prison infraction, even if high, justifies the legal intervention being considered (e.g., lengthened imprisonment).

Additionally, for reasons suggested above, the decision-maker should keep in mind that using low-level arrests as an outcome measure can produce racially disparate results,[56] and that a sizeable number of arrests do not result in convictions.[57] As a result, an RAI that relies on drug possession or misdemeanor arrests as an outcome variable may rate black defendants as higher risk than white defendants who will engage in similar criminal behavior. While that rating might be "validated" because blacks have been and will be *arrested* for more drug possession and misdemeanor crimes than whites, they in fact have not committed more such crimes, and so the risk differential between whites and blacks will not reflect reality.

Violent offending can also be defined in many ways.[58] Some RAIs include within this term any assault or threat of violence, while others may limit that term to homicide, rape, robbery and aggravated assault, or to crimes involving victim injury. Again, the legal decision-maker should find out the outcome variable for the RAI in question and take that information into account when making a decision about whether a person poses sufficient risk to warrant different legal treatment. It has been argued that, where significant deprivations of liberty are involved,[59] violent crime is the appropriate outcome focus, a position that can significantly reduce the number of people considered high risk.[60]

3. *Is the **time frame** associated with the predicted conduct consistent with the legal decision at issue?*

In many legal settings, the legal decision-maker is attempting to forecast reoffending within a limited period of time. For instance, a judge may be sentencing an offender for a crime with a maximum sentence of one year, or a parole board may be considering the risk posed by an individual between the time of its decision and the next parole review period a year or two hence. The risk information provided by some RAIs may not be relevant in such settings. For instance, while the COMPAS provides probability estimates for a period of one to two years, the HCR-20's time frame is two years, and the VRAG's time frame is seven years. Legal decision-makers should take this type of durational

criterion into account in making decisions based on the results of an RAI.

4. *Does the RAI provide information about the effect of different **types of interventions** on the individual's risk categorization?*

An individual's risk of reoffending can vary enormously depending on the law's response to the risk. While imprisonment might substantially reduce risk, it may not be any more effective at doing so than placement in an effective substance abuse treatment program in the community, a vocational training program in a halfway house, or a job-release program coupled with an ankle monitor. Some RAIs, like the HCR-20 and the LSCMI, attempt to provide information about an individual's criminogenic needs that can help make such determinations. RAIs that focus solely on protective factors such as coping skills and supportive social networks are also available to legal decision-makers.[61] These RAIs can be very helpful in fashioning dispositions, although legal decision-makers should make sure to inquire about and attempt to take into account whether a given risk or needs factors is "causal," in the sense that changing or responding to it will *reduce risk* (as opposed to merely "improve" a person). For instance, as noted earlier, while effective substance abuse can have a significant impact on risk, standard psychotherapies are much less likely to do so.

When the legal-decisionmaker is attempting to fashion a disposition to reduce risk, also possibly relevant is the considerable research that finds that incarceration can be criminogenic for a wide range of offenders, given the resulting loss of connection with family, jobs and community and the development of criminal networks that results from imprisonment.[62] A significant amount of research indicates that risk-reducing treatment is often most effective when it takes place in the community.[63] In part for these types of reasons, several commentators have proposed that, for individuals considered low risk, there be a presumption in favor of alternatives to prison or outright release.[64]

Accuracy/Validity

Legal decision-makers should seek assurances that RAIs on which they rely are valid (i.e., do what they purport to do) in the following senses: (1) calibration; (2) discrimination between high and low risk individuals; (3) local validation; and (4) inter-rater reliability. Further, they should seek assurance that (5) these measures of validity are current. Such assurances may come from outside entities (e.g., state agencies, independent peer reviewers, appellate courts); validity determinations do not necessarily need to be made by the legal decision-maker in individual cases. However, legal decision-makers should have an obligation to ensure that, in individual cases: (6) evaluators are trained in the use of the RAI and reliably score the RAI; and (7) any adjustments to the RAI score have a substantial empirical or legal basis.

5. *What evidence is provided about the RAI's calibration?*

There are two primary means of evaluating the accuracy or predictive validity of an algorithm: "calibration" and "discriminant" validity.[65] Calibration, discussed here, measures the extent to which a positive finding (that a person will recidivate) is correct, and the extent to which a negative finding (that a person will not recidivate) is correct. Discriminant validity, discussed next, measures the extent to which an RAI has differentiated between recidivists and non-recidivists, and thus provides a measure of how much better than chance an RAI performs.

Among tool developers, the positive predictive value (PPV) and negative predictive value (NPV) are often thought to be the most

relevant measures of calibration.[66] The question the PPV answers is how often a tool's prediction that someone will recidivate is correct, and the NPV indicates how often a tool's prediction that a person will not recidivate is correct. Usually the PPV is calculated with a confidence interval of 95% (meaning that the range provided by the interval has a 95% chance of correctly identifying the group's risk level).

For instance, using a sample dataset from the COMPAS which divided offenders into high, medium and low risk groups, Melissa Hamilton calculated that the PPV for those designated as a high risk of violent reoffending is 49%, with a confidence interval range of 43% to 55%;[67] in other words, there is a 95% chance that a high risk score on the COMPAS is associated with a 43 to 55% probability of recidivating.[*] Hamilton also found that the NPV of the medium and low risk group examined together was 86%, with a confidence interval range of 85% to 87%, meaning that there is a 95% chance that a medium or low risk score is associated with an 85 to 87% chance of not recidivating.[68] When, instead, Hamilton lumped together the medium and high-risk groups, separate from the low risk group, the PPV with respect to whether the first group recidivated fell to 31% (29% to 34%) and the NPV with respect to whether the second group did not recidivate went up slightly to 89% (88% to 91%).[69]

This example illustrates two important points. First, the PPV and NPV can be manipulated by changing the cut-point (for instance, as Hamilton did, from the high category to the high-medium category combined). Second, to arrive at PPV and NPV, one has to assume that the cut-off, whatever it is, is equivalent to a prediction that the individuals in those groups *will* recidivate. In fact, that is not what most RAIs claim to be doing, or at least not what they should claim to be doing. RAIs cannot identify who will recidivate and who will not recidivate. Rather, as noted above, they can only associate a particular score or category with a *group probability* of reoffending.

Thus, while PPV can help figure out how well an RAI is calibrated, a more legally relevant measure of calibration is what could be called the *category* positive predictive value, or category base rate (CBR),[70] which answers the following three-part question: (1) does the RAI associate a person with a category (a score, decile or risk group); (2) if so, what percentage of people in the category does it predict will recidivate in the way defined by the relevant law; and (3) to what extent is that percentage correct, as measured by validation studies and confidence intervals? An example of this type of RAI is the VRAG. That tool assigns people with certain scores to nine bins associated with specific ascending probability ranges of recidivating.

From a legal perspective, the CBR is the key measure of validity. RAIs that fail to provide CBRs obscure the normative decision about whether the risk an individual's group poses warrants some type of legal intervention. Unfortunately, as discussed earlier, some RAIs do not use scores or other categories, or do so, but associate them only with deciles or undefined high, medium and low risk labels rather than with probabilities.

In some cases, a policy-making body approves the cut-off scores, based on an assessment of CBR and other measures of validation. For instance, in Virginia, the legislature directed the state sentencing commission to develop an instrument that flagged the lowest 25% in terms of risk; the resulting instrument recommends alternatives to imprisonment for that group.[71] Thus, the

[*] It should be noted that the confidence interval range could skew toward over or under prediction, which suggests that the reported proportion should be the focus of the risk assessment.

Virginia legislature mandated the legally relevant cut-off score that legal decision-makers should consider. However, if this type of upper-level policy decision has not taken place, the legal decision-maker should always endeavor to ascertain the CBR for an individual's risk category.

6. *What evidence is provided about the ability of the RAI to differentiate between recidivists and non-recidivists?*

An RAI can have very good calibration—that is, its probability forecasts may be borne out—but can still be of questionable predictive validity. For instance, when a given sample has a low base rate of reoffending (say 5%), an RAI might do no better than "a naïve classifier that predicts that no one recidivates,"[72] since such a classifier would be right 95% of the time. Yet the resulting classifications would be useless to decision-makers who want to separate those who are high risk from those who are not. In recognition of this problem, another measure of predictive validity, one that examines the extent to which an RAI can discriminate between high and low risk individuals independently of how accurately it identifies absolute risk levels, should be sought and provided.

The most popular such measure is called the AUC, for Area Under the Curve. The "curve" is created by plotting, for each point total or cut point (e.g., decile) of the instrument, the true positive rate or "sensitivity" (the rate at which those who recidivated received that score) against the true negative rate or "specificity" (the rate at which those who did not recidivate received that score).[73] If the curve created by this plot is a diagonal, the AUC is .5, and represents a finding that the tool does no better than chance at designating those who recidivated from those who did not, whereas a curve that looks like an "r" would indicate a perfect ability to do so.

Virtually all RAIs have AUCs above .5, with most developers reporting AUCs of .6 to .85.[74] The latter figure means that, 85% of the time, a randomly selected recidivist receives a higher score on the tool than a randomly selected non-recidivist. Social scientists have designated a .56 AUC as a small effect size, .64 as a moderate effect size, and .71 as a large effect size.[75] Ultimately, however, the discriminant validity threshold should be legally determined.

As an evidentiary matter, the AUC value can be thought of as a measure of relevance, which under the rules of every state requires that proffered evidence have a "tendency to make the existence of any fact that is of consequence to the determination of the action [here reoffending] more probable or less probable than it would be without the evidence."[76] Under this definition, the results of an RAI—even one with a useful PPV—that has an AUC of .5 or close to it might be considered irrelevant. Such an instrument's designation of someone as high risk or low risk is no more accurate than a *random* designation of individuals as either high or low risk (the proverbial coin flip). An adequate AUC threshold might be particularly important to enforce in settings known to involve populations with low base rates of offending, such as offenders convicted of capital murder who will clearly be incarcerated for the foreseeable future.

7. *Has the RAI been validated on the jurisdiction's population and with respect to important sub-groupings?*

An RAI is usually developed on half a sample (the development sample) and then tested on the other half of the sample (the validation sample) to see how well it performs. But to be optimally useful in jurisdictions outside of the one in which the RAI was developed, the RAI should perform well in those jurisdictions as well.[77] One of the initial criticisms of the VRAG is that it was validated on a sample of incarcerated

white men with mental disability in Canada. Until it was tested on diverse populations in the United States, its validity with respect to those populations was suspect.[78] Likewise, RAI validation should be attentive to the type of population that will be evaluated. An RAI validated on a population of sex offenders, using sexual assault as the outcome variable, will be of limited use for assessing the risk of other types of offenders.[79] An instrument validated on a sample in an urban area might not work well in rural areas with lower crime rates and fewer arrest and convictions.[80]

Finally, calibration and discriminant validity should also be tested with respect to important demographic groups within the jurisdiction, to ensure predictive validity is maintained for those groups. Research for many of the most widely used instruments shows that calibration and discriminant validity do not vary significantly for most groups.[81] But they can. For instance, Hamilton demonstrates that, for the males in the COMPAS sample described in query #5, the CBR goes up for each of the ten deciles, meaning that the predicted probability of recidivism and the actual rate of recidivism for each decile are fairly closely aligned. However, for the female part of the sample, the CBR does not align well with actual recidivism from the fourth decile on, and drops precipitously at the seventh through ten deciles (with females in the higher risk categories recidivating at a much lower rate than males).[82]

AUC values can also differ depending on ethnicity. Using the COMPAS sample, Hamilton calculated that the AUC for blacks (at the 95% confidence level) is .71 (.68 to .74), for whites .68 (64 to .73) and for Hispanics .64 (.55 to .73).[83] For Hispanics, the AUC on the COMPAS comes perilously close to dipping below an acceptable level of relevance. In recognition of this type of problem, in 2018, the Canadian Supreme Court prohibited use of a risk assessment instrument that was validated on the majority population to assess the risk of a person from an indigenous group.[84]

In short, to be valid, an RAI should have similar positive predictive values for each risk classification across as many major demographic groups as possible, and have a similar AUC for each demographic group as well.[85] This may require separate instruments (with different risk factors) for some groups, such as people of color, women, and so on (the legal implications of which are discussed in connection with query #13 below). Achieving this degree of validation can be difficult, because it requires large enough samples to arrive at statistically useful findings in each of these types of categories. Nonetheless, from an empirical perspective, serious effort should be made to validate the RAI on a population as similar as possible to the offender's reference group. Otherwise, even a tool which, on its face, has satisfactory calibration and discriminant accuracy verges on being irrelevant, a possibility that several courts have noted.[86]

8. *What evidence is there concerning inter-rater reliability of evaluators who use the RAI?*

To a social scientist, reliability means repeatability, the ability to produce similar results under similar circumstances. An RAI that has good calibration, discriminant, and external validity may still not produce valid results if it cannot be administered reliably. Some RAIs, like the HCR-20, rely on numerous "soft variables," such as "lack of insight," "negative attitudes," and "impulsivity." These types of variables are subject to many interpretations. Even more objective instruments, such as the VRAG, include items that can suffer from poor inter-rater reliability, such as the individual's score on the Psychopathy Checklist-Revised and his or her psychiatric diagnosis. Unless evaluators measure these types of factors in a way that is consistent

with the how they were conceptualized by the tool's developers, the chance that the same individual will receive different scores from different evaluators or that similar individuals will receive different scores from the same evaluator is significant.[87]

A survey of 53 studies in 2013 found that only two reported inter-rater reliability.[88] Ideally, this information would be reported for every RAI. An agreement ratio of 80% among raters is considered very good.[89]

9. *How current is the evidence supporting the RAI's validity?*

All of the measures of accuracy discussed to this point can alter for a particular instrument if, for instance, the population on which the RAI was normed changes substantially, the jurisdiction's crime, arrest or conviction rates go up or down significantly, or the jurisdiction begins implementing innovative alternatives to prison that can reduce risk.[90] Because the potential for offending is affected by these types of factors, a person rated high risk on an outdated instrument may actually belong to a group that is low risk, or vice versa. For that reason, RAIs should be re-validated periodically. Many of the most prominent instruments have been revised based on new data. For instance, Virginia's sentencing RAIs have been re-validated twice since 2001.[91]

Other measures of current validity include case auditing and peer review.[92] Tools tested in the field rarely do as well as they did during validation. Periodic auditing of how they are implemented by evaluators is one way of ensuring RAIs are performing adequately.[93] Virginia, for instance, provides annual reports on its RAIs.[94]

The best way to ensure current validity is through peer review, carried out by researchers who did not develop the instrument. University-based experts are often best situated to carry out this type of review. CBRs, AUCs and reliability data that have been replicated by such experts, ideally in the same setting in which the RAI is being used, instill confidence that the RAI is reliable in both the legal and scientific sense.

10. *Has the person or persons using the RAI been trained in its use, are the evaluator's conclusions about the individual's risk factors accurate, and is the risk score calculated properly?*

Federal Rule of Evidence 702 states that the basis of expert testimony must derive from reliable methods and principles, "reliably applied."[95] While evidentiary rules like this one are not formally applicable to post-conviction proceedings in many jurisdictions, the basic principle should not be ignored where deprivations of liberty are at stake. Applied to risk assessments, this language makes it incumbent on the decision-maker to ensure that the evaluator has been trained in using the particular RAI, and has relied on trustworthy information sources in gathering relevant data.[96] It also requires some assessment of whether the scoring and assignment of the individual to a particular risk group was carried out in a competent manner.

11. *If the legal decision-maker is contemplating "adjusting" the risk score, are there valid empirical or practical reasons for doing so that have not already been considered by the developers of the RAI?*

One survey of judges who use an RAI found that even those who were highly favorable toward the instrument "were still inclined to consider recommendations in the context of their own judicial intuition and experience, and would request information that was not included in the risk assessment instrument when they deemed this to be necessary."[97] The conclusion of one judge is typical: "It's important to

understand that it's just a tool and that judges are the definitive answer."[98] The judicial decisions that have analyzed the use of RAIs at sentencing have likewise emphasized that the results of an RAI is but one factor to consider and should not be dispositive.[99]

If these statements are merely meant to stress that sentencing judges should always consider other purposes of punishment besides risk and incapacitation, they are unremarkable. But if these statements are asserting that, even when focused solely on the question of risk, legal decision-makers should feel free to second-guess the results of a well-validated RAI, they should be tempered with the knowledge that such adjustments can easily reduce accuracy. Consistent with the findings about RAIs' incremental validity compared to decisions made by lay factfinders, evaluator and judicial adjustments usually do not improve on the actuarial assessment. In fact, several studies find that professional "overrides" of an RAI's risk estimate, whether by judges, probation officers or other correctional professionals, *decrease* accuracy in predicting offending.[100] For example, based on a sample of 3,646 offenders, Guay & Parent found that adjustments to an RAI result made by probation officers were significantly less accurate than the unadjusted RAI.[101] A study by Schmidt et al. found that professional overrides decreased predictive validity and usually increased risk level.[102] The most recent study likewise found that overrides typically result in an "upward reshuffling" of risk, and a loss of predictive accuracy.[103]

There are likely several explanations for these types of findings. Adjustments may be based on unverified speculation about the traits that might affect risk, a belief that "special circumstances" (e.g., contriteness or surliness) warrant ignoring the risk score, or simple mistrust of quantified decision-making.[104] Or they may stem from extraneous considerations. Erroneous adjustments upward, to a higher risk label, may be influenced by knowledge that a false negative decision, which results in release, is much more likely to be discovered than a false positive decision that results in incarceration; moreover, of course, the latter types of errors are much more likely to have professional and societal consequences.[105] In contrast, erroneous decisions downward may reflect concern about whether sufficient treatment resources are available in prison; if not, judges have been known to opt for a prison alternative that appears to better serve the individual (although here the override is more likely to be warranted, if the RAI has not taken the community treatment into account).[106] More generally, evaluators, judges, and parole board members might dislike the idea of having their decisions dictated by a table; as one Virginia judge put it, "I don't do voodoo."[107] Unfortunately, because these adjustments to the RAI result are at best based on untested assumptions derived from experience, they may not only be wrong but also infected by racial and other biases, a possibility some courts have noted.[108]

The dangers of adjusting a risk level in the absence of supporting research is particularly high with upward adjustments.[109] Most RAIs consist primarily of risk factors, not protective factors, so the reason for rating a person as a higher risk than the RAI indicates can easily be something that has already been taken into account. One of the common mistakes in this regard is to "double-count" criminal history. For instance, a judge might decide that even though an RAI indicates an offender poses a low risk, the sentence should be enhanced because the offender has committed two prior offenses; research in Virginia indicates that is precisely what happens there.[110] Since every RAI already incorporates criminal history into its algorithm, this assessment will almost certainly be erroneous, and thus decrease the

reliability/validity of the risk assessment enterprise.

In contrast, relevant protective factors—that is, again, traits that are correlated with lower risk—are less likely to have been considered during the development of an RAI. Further, certain types of interventions can reduce risk. If an individual is able to produce research showing the presence of protective factors not considered in the RAI's development, or that a particular intervention that fits his or her criminogenic needs is available, a downward adjustment or some alternative to prison may be indicated. In practice, RAI-overrides that can be justified on solid evidence will normally be in the downward direction.

Justice/Fairness

The legal decision-maker should evaluate the fairness of RAIs by ensuring that: (1) disparate impacts on major demographic groups can be justified on empirical grounds; (2) the instrument does not rely on risk factors that are barred from consideration by the jurisdiction; (3) the risk factors and how they were combined are accessible to all parties, including the subject of the risk assessment; and (4) the subject of the risk assessment is provided an explanation of the risk factors and the empirical logic behind the RAI, is able to contest the RAI analysis of risk factors and its results, and can present evidence of protective factors.

12. *Are there data that provide information about the impact of the RAI on different racial groups and other important groups?*

Race and sex are protected classes under the equal protection clause of the Fourteenth Amendment. An unresolved legal question is whether attempts to avoid unfair impacts on these two groups by taking race or sex into account in constructing an RAI violates equal protection principles. This can occur in at least two ways. First, as discussed above (see query #7), in an attempt to improve accuracy, an RAI may be calibrated for each major ethnic group, an adjustment that might be called "race-conscious calibration." Second, in an effort to assure what some commentators have called "classification parity," false positive rates could be adjusted. For instance, if an RAI produces higher false positive rates for blacks or other ethnic groups than for whites, risk categories could be adjusted so that the risk scores for people of color are lowered or the risk scores for whites are raised.

Both race-conscious calibration and classification parity explicitly use race as a discriminator, a fact that implicates the equal protection clause.[111] There is a key difference between the two types of adjustments, however. Unlike classification parity—which changes an accurate conclusion about the statistical likelihood of recidivism to achieve its version of fairness—race-conscious calibration serves the important state interests of protecting the public and avoiding unnecessary incarceration, by rectifying the impact of discriminatory practices that unfairly raise one's risk score. As between the two, race conscious calibration is more likely to survive an equal protection challenge.

While no RAI explicitly incorporates race as a risk factor (and, according to the Supreme Court, would clearly be unconstitutional if it did so[112]), sex—specifically, maleness—is a risk factor in several RAIs. Developers have realized that an RAI that is well-calibrated for men may not be well-calibrated for women; women do not recidivate as often as men, apparently even when they are otherwise associated with identical risk factors. From an empirical point of view, that situation calls for an RAI validated on a female population. But, again, such an explicit use of a protected class could be seen to be

violative of the equal protection clause. For instance, in *Craig v. Boren*,[113] the Supreme Court struck down a law that allowed women to buy alcohol at age 18 while prohibiting alcohol sales to males until they are 21, despite evidence that men have higher rates of drunk driving.

Nonetheless, in *Wisconsin v. Loomis*[114] the Wisconsin Supreme Court suggested that discriminating on the basis of sex is permissible if it validly helps distinguish between males and females in terms of risk. Loomis' sentence had been enhanced using the COMPAS, which specifically took gender into account; Loomis argued that this disposition violated due process. Although as a result of this framing, the Wisconsin court did not explicitly address the equal protection issue, it did state, in the course of rejecting Loomis' claim, that "it appears that any risk assessment tool which fails to differentiate between men and women will misclassify both genders."[115] In essence, the court was saying, because of its enhancement to accuracy, incorporating gender was a narrowly tailored means of meeting the state's interest in preventing harm to the public in a cost-efficient manner.

Of the factors typically found in RAIs, only race and sex trigger Fourteenth Amendment protection and thus require more than a rational basis for their use under current law. Nonetheless, it has been argued that the Fourteenth Amendment also bars RAIs from using poverty or proxies for it (e.g., unemployment, location, or house ownership),[116] based primarily on the Supreme Court's pronouncement in *Bearden v. Georgia*[117] that revoking parole for an offender who has failed to pay a fine "would be little more than punishing him for his poverty," and "is contrary to fundamental fairness."[118] However, no court has interpreted *Bearden* to mean that factors *related* to poverty are anathema in assessing either risk or punishment generally;[119] *Bearden* itself stated that "a sentencing court can consider a defendant's employment history and financial resources in setting an initial punishment,"[120] and emphasized that the only sentencing practice it was barring was the use of poverty "as the sole justification for imprisonment,"[121] which no risk assessment instrument does.

This does not mean that any wealth-related risk factor is fair game. Following equal protection's tiered analysis, the use of factors other than race and sex—such as age, employment status, home life as a child, diagnosis or marital status—must still have a rational basis. Because of their minimal predictive value, for instance, employment and marital status were eventually dropped from Virginia's NVRA.[122]

Finally, an RAI might give rise to a disparate impact claim rather than a disparate treatment claim. Again, however, if an RAI uses neither race nor sex as a risk factor, but only produces results that have a disparate racial or gender impact, then formal classification is not occurring, and use of the RAI is permissible if there is any rational basis for doing so, unless a discriminatory purpose can be shown. While the Supreme Court has not always required serious animus in its disparate impact cases, it has tended to require strong proof of discriminatory purpose in criminal cases.[123] In any event, developers of RAIs are not likely to have harbored or intended to implement animus toward any given racial group, and in fact presumably want to avoid disproportionate outcomes.[124] Thus, a disparate impact argument against RAIs is unlikely to prevail.

13. *Do any of the risk factors in the RAI rely on traits that the jurisdiction's law prohibits legal decision-makers from considering at sentencing?*

Independent of equal protection concerns are enactments that specifically prohibit reliance on certain types of factors in punishment decisions. For instance, Ohio law states that "A court that imposes a sentence upon an offender for a felony shall not base the sentence upon the race, ethnic background, gender, or religion of the offender,"[125] and Tennessee law provides that "Sentencing should exclude all considerations respecting race, gender, creed, religion, national origin, and social status of the individual."[126] Thus, for instance, both statutes could be construed to bar sex as a risk factor, regardless of the effect of that prohibition on accuracy. Furthermore, Tennessee's prohibition on punishment that considers the "social status of the individual" might bar consideration of any factors having to do with income, employment, marital status, and the like. It is a matter of state law whether a sentence that relies in whole or part on an RAI that incorporates such factors is "based" on those factors or explicitly "considers" them. If so, use of RAIs that include such factors may be impermissible.

14. *Are the RAI's risk factors and the weights assigned to them accessible to peer reviewers, evaluators, legal decision-makers and the subjects of the risk assessment?*

Some RAIs are developed by private companies that claim trade secret protection over the algorithm. For instance, the company that produces the COMPAS claims its algorithm and the weights it assigns risk factors are protected, a claim that the Wisconsin Supreme Court upheld.[127] In such a situation, the identity and importance of risk factors can be difficult to discern. For instance, sophisticated reverse engineering was required to figure out that, while the COMPAS contains over 100 items, over half of the risk score is attributable to a single factor, the offender's age.[128] Even purportedly publicly developed instruments can be less than transparent. Congress required that the RAI developed under the federal First Step Act, the PATTERN, be made public,[129] but did not require that the validation procedure that led to development of the instrument nor the data underlying it be disclosed. When asked for more information, the authors of PATTERN stated that state law-driven privacy concerns prevented release even of anonymized versions of the data to outside researchers.[130] A number of states have responded to similar requests in the same fashion.[131]

The integration of sophisticated machine learning into RAI construction could make RAIs even more opaque, since under some versions of that technique the weights assigned risk factors and even the identity of those factors are inaccessible to humans.[132] Furthermore, even if the black box can be opened, serious interpretation problems can arise. More specifically, as Andrew Selbst and Simon Barocas note, some versions of machine learning can be either "inscrutable"—meaning that even when a model is available for direct inspection it may "defy understanding"—or "non-intuitive"—meaning that even where a model is understandable it may "rest on apparent statistical relationships that defy intuition."[133]

Even if it turns out that advanced RAIs are demonstrably more accurate than simpler versions (which is unlikely[134]), and putting aside whether actuarial instruments need to be intuitively understandable, algorithms that are "inscrutable" are problematic.[135] Neither equal protection nor statutory analysis of the type just described can take place unless the legal decision-maker can discern the risk factors in an RAI. The accuracy of the probabilities and other results reached by an RAI cannot be confirmed unless the underlying data and the empirical analysis using it can be evaluated by others. If decision-makers want to avoid "adjustments" of

a risk assessment based on factors already considered in developing the instrument they need to know what those factors are. Perhaps most importantly, without transparency litigants cannot adequately contest the facts leading to the RAI's results.

To aid both independent peer reviewers and those involved in the legal process, developers should provide "a complete description of the design and testing process . . . , [a] list of factors that the tool uses and how it weighs them, [t]he thresholds and data used to determine labels for risk scores, . . . [t]he outcome data used to develop and validate the tool at an aggregate and privacy-protecting level--disclosing breakdown of rearrests by charge, severity of charge, failures to appear, age, race, and gender—[and] clear definitions of what an instrument forecasts and for what time period."[136]

Some caselaw backs up these requirements. In *Gardner v. Florida*,[137] the defendant argued that, before his sentence was imposed, he had a due process right to discover and rebut the contents of his pre-sentence report. The Supreme Court agreed, stating: "Our belief that debate between adversaries is often essential to the truth-seeking function of trials requires us also to recognize the importance of giving counsel an opportunity to comment on facts which may influence the sentencing decision . . . ".[138] While *Gardner* was limited to the death penalty context, the Court came to a similar conclusion in *Roviaro v. United States*,[139] a simple drug case. There, the Court held that the identity of confidential informants must be revealed to the defendant when the informant possesses facts that are highly relevant to the defense. *Roviaro* establishes that even strong claims of a need for secrecy (here protecting an informant) should not prevail when the information is crucial to the case. While *Roviaro* has been given short shrift in more recent lower court decisions,[140] its central rationale has not been abandoned.[141]

Scholars have also made sub-constitutional arguments in favor of open algorithms. Danielle Citron has contended that private companies that seek public money for products that affect public policy should not be able to hide behind trade secret laws,[142] and Rebecca Wexler has noted that companies' concern about algorithmic disclosure giving competitors an advantage or discouraging innovation are overblown, especially if protective orders or *in camera* review requirements are imposed.[143] The opacity problems created by machine learning have received special attention. Most prominently, scholars have argued for a "right to explanation,"[144] a right that the European Union has explicitly recognized in its General Data Privacy Regulation.[145]

15. *Are the subjects of the risk assessment provided an explanation of the empirical basis of the RAI and the opportunity to contest and rebut its results?*

In *State v. Guise*, Judge Appel stated: "[O]ne thing is clear: if the state intends to offer risk assessments for the court to rely upon in sentencing, the defendant has a right to an adequate opportunity to attack it. If the court does not give the defendant an adequate opportunity to attack the statistical evidence, it should not be utilized in sentencing."[146] Yet in many jurisdictions, the adversarial process provided at criminal trials does not exist post-conviction. Even at sentencing, the adjudication process can be very informal, at least outside of capital cases.[147] Defendants are entitled to be present during sentencing, but do not have a right to testify in the normal sense, only a right to "allocution" (a statement by the defendant that can be restricted to a plea for mercy);[148] funding for expert testimony is also minimal

outside of capital cases.[149] In most states and in federal court, neither *Daubert* nor any rule resembling it applies at sentencing.[150] The process due at other post-conviction proceedings is even more minimal.[151] All of this makes relevance, accuracy and fairness challenges difficult.

In the absence of a full-blown adversarial process, other procedural components are crucial. Most importantly, the subject of the risk assessment should be able to raise claims about whether the RAI was reliably administered, in two ways. First, the subject should be able to attack the accuracy of a conclusion that a particular risk factor is present (e.g., the validity of an assumed arrest or conviction, the applicability of a diagnosis, or the failure to complete a program that in fact was not available to the offender). Second, the subject should be able to proffer protective factors that were not considered by the developers of the instrument (e.g., completion of a treatment or educational program, changes in employment status); researchers are beginning to identify a number of such factors.[152] Third, the state's evaluator should provide information about any perceived protective factors that the subject does not identify, especially when subjects do not have access to their own experts. As a supplement to these process rights, the legal decision-maker should provide a written explanation for any decision based on risk, one that should be particularly detailed if an adjustment to the RAI results occurs.

RAIs should also be subject to legislative and administrative review. A number of state legislatures have mandated that sentencing judges or corrections officials use a "validated risk assessment tool,"[153] and in other states the state sentencing commission,[154] the department of corrections,[155] the state courts generally,[156] or, as California has done with respect to pretrial risk assessments, the courts in each jurisdiction,[157] have taken on the task. In the federal First Step Act, Congress directed that (1) the Attorney General develop and release a "risk and needs assessment system" to determine the "recidivism risk of each prisoner" following "an objective and statistically validated method," (2) a panel of researchers approve the instrument, (3) the instrument be annually validated, and (4) Bureau of Prison staff "demonstrate competence in administering the System, including interrater reliability, on a biannual basis."[158] While the analysis of the RAI could be under the auspices of the department of corrections or the sentencing commission, ideally the type of research panel referred to in the Act would consist of experts outside of the department, perhaps at a university, as suggested in the discussion of query #9.

This type of independent review of RAIs would also take the burden off individual legal decision-makers, who might have difficulty answering many of the inquiries outlined in this primer. In fact, such a panel could presumptively resolve many of the relevance, validity, and fairness issues described above. Specifically, while the issues set out in 4 (selecting risk-reducing interventions), 10 (ensuring the evaluator has been adequately trained and has properly scored the RAI), 11 (deciding whether adjustments to the RAI results should occur), 14 (ensuring that the contents of the RAI are accessible), and 15 (ensuring adequate adversarial testing) would have to be handled by judges, parole boards and correctional officials on a case-by-case basis, the remaining issues could be addressed in the first instance by the outside entity.

CONCLUSION—THE NEED FOR A JURISPRUDENCE OF RISK

Until recently, the post-conviction use of risk assessment instruments has received little attention from the legal community, at least in

comparison to the vast literature focused on defining crimes, defenses, and the amount of punishment that particular types of offenders "deserve." Without a jurisprudence of risk, judges and other legal decision-makers have very little guidance on which risk assessment instruments, if any, are worthy of consideration, how to evaluate their results, and how much weight to give those results. Researchers who develop these instruments do not have a clear idea of the outcome measures that the law considers relevant, the types of risk factors they may or may not consider, or the psychometric properties that ensure courts will rely on the instrument they develop. In the meantime, to the extent risk assessment influences the post-conviction process, the fate of offenders, the safety of the public, and even the pace of incarceration rates are subject to a hodgepodge of inchoate views about the impact an offender's risk should have on disposition. In addition to providing useful information to legal decision-makers about risk assessment instruments, a central aim of this primer is to encourage further development of this jurisprudence of risk.

REFERENCES

[1] John Monahan & Jennifer Skeem, Clinical and Actuarial Predictions of Violence: Scientific Status, in Modern Scientific Evidence: The Law and Science of Expert Testimony 179-180 § 9.11 (2019-2020).

[2] Kevin Douglas et al., Historical-Clinical-Risk Management 20, Version 3 (HCR020 v.3): Development and Overview, 13 Int'l J. Forensic Mental Health 93 (2014).

[3] Northpointe, COMPAS Risk and Needs Assessment System, http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf (accessed, July, 2020).

[4] James Bonta & D.A Andrews, Risk-Needs-Responsibility Model for Offender Assessment and Rehabilitation (2007), https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/rsk-nd-rspnsvty/rsk-nd-rspnsvty-eng.pdf

[5] Grant Harris et al., Violent Offenders: Appraising and Managing Risk (3d ed., 2015).

[6] Richard Kern & Meredith Farrar-Owens, Sentencing Guidelines with Integrated Offender Risk Assessment, 16 Fed. Sent. Rep. 165 (2004)

[7] Andrew J. R. Harris, Amy Phenix & Karl M. Williams, STATIC-99 Clearinghouse, http://www.static99.org/ (accessed, July, 2020).

[8] Daryl Kroner, Jeremy Mills & John Reddon, A Coffe Can, Factor Analysis, and Prediction of Antisocial Behavior: The Structure of Criminal Risk, 28 Int'l J. L. & Psychiatry 360 (2005) (finding that most RAIs tap for overlapping dimensions of risk: (1) criminal history; (2) irresponsible lifestyle (e.g., poor engagement in school/work; (3) psychopathy and criminal attitudes (e.g., feelings of entitlement); and (4) substance abuse-related problems).

[9] See generally, Jennifer Skeem & John Monahan, "Risk," "Needs" and "Evidence" in Implementing the First Step Act, 38 Beh. Sci. & L. 287 (2020).

[10] In re Linehan, 518 N.W2d 609, 616 (Minn. 1999) (Coyne, J., dissenting). See also See Porter v. Comm, 661 S.E.2d 415 (Va. 2008) (excluding defense testimony on this ground); United States v. Taylor, 583 F.Supp.2d 923 (E.D. Tenn. 2008) (excluding defense testimony that "invites the jury to make decisions based upon group characteristics and assumptions); Rhodes v. State, 896 N.Ed.2d 1193 (Ind.Ct.App. 2008) (holding that using an LSI-R (LSCMI) score as an aggravating factor at sentencing was impermissible).

[11] See David Faigman, John Monahan & Christopher Slobogin, Group to Individual (G2i) Inference in Scientific Expert Testimony, 81 U. Chi. L. Rev. 417 (2014).

[12] Peter B. Imrey & A. Philip Dawid, A Commentary on Statistical Assessment of Violence Recidivism Risk, 2 Stat. & Pub. Pol'y 1, 1 (2015)

[13] R. Karl Hanson & Philip D. Howard, Individual Confidence Intervals Do Not Inform Decision-Makers About the Accuracy of Risk Assessment Evaluations, 34 Law & Hum. Behav. 275, 277 (2010).

[14] Barbara Underwood, Law and the Crystal Ball: Predicting Behavior with Statistical Inference and Individualized Judgment, 88 Yale L.J. 1408, 1427 (1979).

[15] See Erica Beecher-Monas, Lost in Translation: Statistical Inference in Court, 46 Ariz. St. L.J. 1057, 1092 (2014) (calling this type of formulation a "solution to the G2i problem in the courts" because "the instrument does not say . . . that the tested individual is 26% likely to recidivate; rather it says [h]e is merely part of a group with an average recidivism rate of 26%. Some will recidivate more, some less. The inference about the tested individual's likelihood of recidivism is left to the factfinder.").

[16] Julia Dressel & Hany Farid, The Accuracy, Fairness, and Limits of Predicting Recidivism, 4 Sci. Adv. EAAO5580 (Jan. 18, 2018), available at https://advances.sciencemag.org/content/4/1/eaao5580.

[17] Id. at 3.

[18] Sharad Goel et al., The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment (2019).

[19] Zhiyuan "Jerry" Lin et al., The Limits of Human Predictions of Recidivism, 6 Science Advances EAAZ-0652, at 2-4 (Feb. 20, 2020), https://advances.sciencemag.org/content/6/7/eaaz0652.

[20] Id. at 5 ("we also found that algorithms tended to outperform humans in settings where decision-makers have access to extensive information and do not receive immediate feedback and base rates are far from balanced, features of many real-world scenarios").

[21] See Sarah J. Desmarais, Kiersten L. Johnson & Jay P. Singh, Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings, 13 Psychol. Servs. 206, 206 (2016) available at https://doi.org/10.1002/9781119184256.ch1 ("There is overwhelming evidence that risk assessments completed using structured approaches produce estimates that are more reliable and more accurate than unstructured risk assessments."); R. Karl Hanson & Kelly E. Morton-Bourgon, The Accuracy of Recidivism Risk Assessments for Sexual Offenders: A Meta-analysis of 118 Prediction Studies, 21 Psychological Assessment 1, 6 (2009) ("For the prediction of sexual or violent recidivism, the actuarial measures designed for violent recidivism were superior to any of the other methods."); William M. Grove et al., Clinical Versus Mechanical Prediction: A Meta-analysis, 12 Psychological Assessment 19 (2000) (finding that actuarial predictions are about 10% more accurate than clinical predictions). See also Ben Green & Yiling Chen, Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments (2019) (unpublished manuscript), https://scholar.harvard. edu/files/19-fat.pdf [https://perma.cc/Q8QA-AHHL] (presenting the results

of an experimental study in which human subjects "underperformed the risk assessment even when presented with its predictions").

[22] Stefania Ægisdóttir et al., The Meta-analysis of Clinical Judgment Project: Fifty-six Years of Accumulated Research on Clinical Versus Statistical Prediction, 34 The Counseling Psychologist 341, 368 (2006) (out of 1,000 predictions, statistical predictions of violence accurately identify 90 more violent clients than do clinical predictions).

[23] Id.

[24] Jongbin Jung et al., Simple Rules for Complex Decisions (2017), available at https://arxiv.org/pdf/1702.04690 (finding that when compared to judges in a pretrial setting a simple algorithm that only looked at two factors of arrestees, age and previous failures to appear, "consistently outperform[ed] the human decision-makers"); Jon Kleinberg et al., Human Decisions and Machine Predictions, 133 The Quarterly Journal of Economics 237, 241 (2017) (similar finding); Thomas H. Cohen, Bailey Pendergast & Scott W. VanBenschoten, Examining Overrides of Risk Classifications for Offenders on Federal Supervision, 80 Fed. Probation, 13, 20-21 (2016) (finding that overrides by probation officers were "almost all . . . an upward adjustment" and that these overrides "demonstrated a weaker correlation between the adjusted risk levels and recidivism compared to the original risk levels"); Hanson & Morgan-Bourgon, supra note 21; J. Stephen Wormith, Sara M. Hogg & Lina Guzzo, Wormith, J. S., Hogg, S., & Guzzo, L., The Predictive Validity of a General Risk/needs Assessment Inventory on Sexual Offender Recidivism and an Exploration of the Professional Override, 39 Criminal Justice & Behavior 1511-1538 (2012) (finding that a professional override by probation officers, psychologists, or social workers on the result suggested by a risk assessment instrument designed to test sexual offense recidivism "led to a slight, but systematic, deterioration in the predictive validity"); Kathleen Spencer Gore, Adjusted Actuarial Assessment of Sex Offenders: The Impact of Clinical Overrides on Predictive Accuracy Gain (2007), dissertation, https://lib.dr.iastate.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=16536&context=rtd (finding that psychologist overrides on the results produced by an instrument designed to predict recidivism for sex offenders "failed to even nominally exceed the [instrument] in terms of overall predictive accuracy").. See generally, Stephen D. Gottfredson & Laura J. Moriarty, Clinical Versus Actuarial Judgments in Criminal Justice Decisions: Should One Replace the Other?, 70 FED. PROB. 15, 15 (2006) (stating that over-reliance on human judgment may undermine the accuracy of risk assessment, because probation and parole officers may "concentrate on information that is demonstrably not predictive of offender behavioral outcomes.").

[25] Paul Meehl, Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence 24-28 (1954) (discussing the "broken leg" exception to actuarial prediction, recognizing clear cases where it would be foolish to adhere to the actuarial result, such as when t person predicted to be imminently dangerous has a broken leg that the actuarial instrument does not take into account).

[26] Megan Stevenson & Sandra G. Mayson, The Scale of Misdemeanor Justice, 98 B.U. L. Rev. 731, 769-770 (2018).

[27] See Sonya B. Starr, Evidence-Based Sentencing and the Scientific Rationalization of Discrimination, 66 Stan. L. Rev. 803, 805 (2014); State v. Jennings, 2014-Ohio-2307 ¶ 24 (Ohio Ct. App. 2014) (unsuccessful challenge to use of neighborhood as a risk factor).

[28] See Sandra Mayson, Bias In, Bias Out, 128 Yale L.J. 2218, 2265-2266 (2019).

[29] Jeff Larson et al., How We Analyzed the COMPAS Recidivism Algorithm, ProPublica (May 23, 2016), available at https://www.propublica.org/article/ how-we-analyzed-the-compas-recidivism-algorithm ("Across every risk category, black defendants recidivated at higher rates.")

[30] Sam Corbett-Davies et al., What Makes an Algorithm Fair? Medium (Oct. 28, 2016), https://medium.com/soal-food/what-makes-an-algorithm-fair-6ad64d75dd0c ("If the recidivism rate for white and black defendants is the same within each risk category, and if black defendants have a higher overall recidivism rate, then a greater share of black defendants will be classified as high risk. And if a greater share of black defendants are classified as high risk, then . . . a greater share of black defendants *who do not reoffend* will also be classified as high risk.").

[31] Mayson, supra note 28, at 2271 ("it may not even be possible to equalize both error rates at once. An effort to equalize false-positive rates may widen the disparity in false-negative rates, or vice versa").

[32] Washington v. Davis, 426 U.S. 229, 239 (1976) ("[O]ur cases have not embraced the proposition that a law or other official act, without regard to whether it reflects a racially discriminatory purpose, is unconstitutional solely because it has a racially disproportionate impact.").

[33] In fact, the Supreme Court has indicated that protection of the public is not just a rational basis for risk assessment, it is a compelling one. United States v. Salerno, 481 U.S. 739, 745 (1987) (stating, in the pretrial detention context, that the federal government has "compelling interests in public safety"); Schall v. Martin, 467 U.S. 253, 264 (1984) (same and collecting cases standing for the proposition that "[t]he 'legitimate and compelling state interest' in protecting the community from crime cannot be doubted" and is 'a weighty social objective'").

[34] Ben Green & Yiling Chen, Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments (2019) (unpublished manuscript), https://scholar.harvard.edu/files/19-fat.pdf [https://perma.cc/Q8QA-AHHL];
Jon Kleinberg et al., Human Decisions and Machine Predictions (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_ id=2920398.

[35] As Adam Benforado summarizes this point, "[s]tudies on sentencing have shown that judges are influenced by the race of the defendant . . . . [T]he latest psychological research suggests that much of the skew is not susceptible to conscious control. There is no magic switch to erase a lifetime of exposure to damaging stereotypes that link the concepts of blackness and violence . . . ." Adam Benforado, Can Science Save Justice? 101 Judicature 25, 28 (2017).

[36] Ashesh Rambachan et al., An Economic Approach to Regulating Algorithms, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3597843 (2020) (arguing that if data, training procedure and decision rules are disclosed and monitored for disparate impact, "algorithms strictly reduce the extent of discrimination against protected groups relative to a world in which humans make all the decision").

[37] Sendhil Mullainathan, Biased Algorithms Are Easier to Fix Than Biased People, N.Y. Times (Dec. 6, 2019), https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html. See also, Jon Kleinberg et al., Discrimination in the Age of Algorithms, at 3 (2018) ("[g]etting the proper regulatory system in place . . . has the potential to turn algorithms into a powerful counterweight to human discrimination and a positive force for social good of multiple kinds.").

[38] Michael Tonry, Legal and Ethical Issues in the Prediction of Recidivism, 26 Federal Sentencing Rep. 167, 171 (2014) ("Ascribed characteristics for which individuals bear no responsibility, such as race, ethnicity, gender, and age, should not be included.").

[39] Jurek v. Texas, 428 U.S. 262, 275-76 (1976) (stating that "prediction of future criminal conduct is an essential element in many of the decisions rendered throughout our criminal justice system [mentioning bail, sentencing and parole determinations as examples of such decisions]. . . . The task that a Texas jury must perform in answering the statutory question [about dangerousness of a capital defendant] is thus basically no different from the task performed countless times each day throughout the American system of criminal justice.").

[40] See, e.g., Barefoot v. Estelle, 463 U.S. 880 (1981) (upholding the death sentence of Barefoot based on a finding of dangerousness that depending in part on testimony placing Barefoot in the "most severe category" of sociopaths).

[41] Christopher Slobogin, A Defence of Risk Based Sentencing, in Predictive Sentencing: Normative and Empirical Perspectives 107, 121 (Jan W. de Keijser, Julian V. Roberts & Jesper Ryberg, eds., 2019).

[42] Stephen D. Gottfredson & Michael Gottfredson, Selective Incapacitation?, 478 Annals of the American Acad. Of Pol. & Soc. Sci. 135 (1985).

[43] Jessica M. Eaglin, Beyond Equality: Procedural Constraints for Actuarial Risk Assessments at Sentencing, Cornell L. Rev. (forthcoming, 2020). See generally Malcolm Feeley & Jonathan Simon, The New Penology: Notes on the Emerging Strategy of Corrections and Its Implications, 39 Criminol. 449, 455-57 (1992) ("The new penology . . . is about identifying and managing unruly groups").

[44] Jessica M. Eaglin, Constructing Recidivism Risk, 67 Emory L.J. 59, 88 (2017).

[45] See, e.g., Jodi Viljoen et al., Impact of Risk Assessment Instruments on Rates of Pretrial Detention, Postconviction Placements, and Release: A Systematic Review and Meta-Analysis, Law & Hum. Beh. (2019) (a metareview finding that adoption of RAIs slightly decreased incarceration rates, while neither increasing nor decreasing recidivism). Cf. Christopher T. Lowenkamp, Edward J. Latessa & Alexander M. Holsinger, The Risk Principle in Action: What Have We Learned from 13,676 Offenders and 97 Correctional Programs? 52 Crime & Delinq. 77, 89-90 (2006) (meta-review finding that targeting of offenders identified as high risk by RAIs is more effective at reducing risk); Sigrid Van Wingerden, Johan van Wilsem & Martin Moerings Presentence Reports and Punishment: A Quasi-Experiment Assessing the Effects of Risk-based Pre-sentence Reports on Sentencing, 6 European J. Criminol. 723 (2014) (finding that the availability of risk assessments resulted in "less controlling" and "more diverting" sentencing outcomes in the Netherlands). Research on pretrial RAIs reached similar conclusions. Evan M. Lowder, et al., Effects of Pretrial Risk Assessments on Release Decisions and Misconduct Outcomes Relative to Practice as Usual (2021) ("Adherence to structured guidelines was associated with higher rates of non-financial release, lower rates of financial release, less time in pretrial detention, a greater likelihood of release within three days, and a greater likelihood of pretrial release overall.").

[46] Anne Morrison Piehl, Bert Useem & John J. DiIulio, Jr. Right-Sizing Justice: A Cost-Benefit Analysis of Imprisonment in Three States (1999). See also Bert Useem & Anne Morrison Piehl, Prison State: The Challenge of Mass Incarceration ch. 3, text before note 129 ("concluding, based on three economic cost-benefit studies of five state prison systems, that "[a]lthough it clearly pays on incapacitation grounds alone to incarcerate those at the eightieth percentile [of risk] in all five states, it does not appear to 'pay' to incarcerate those below the median.").

[47] Kevin Reitz, The Compelling Case for Low Violence-risk Preclusion in American Prison Policy, 38 Beh. Sci. & L. 207, 215-216 (2020) (showing that, even if racial disparities persist, an RAI that reduces prison populations by 25% would have a much larger absolute impact on racial minorities than on whites).

[48] Tony Fabelo, Geraldine Nagy, and Seth Prins, A Ten-Step Guide to Transforming Probation Departments to Reduce Recidivism 13-19, 27 (Council of State Governments Justice Center 2011) (discussion of why low-intensity monitoring of low risk offenders can be good policy, because otherwise minor technical violations are discovered that result in unnecessary confinement, activities that result in positive behaviors are disrupted, and treatment failures assume exaggerated importance).

[49] State v. Krol, 344 A.2d 289, 302 (N.J. 1975) ("The determination of dangerousness involves a delicate balancing of society's interest in protection from harmful conduct against the individual's interest in personal liberty and autonomy. This

decision, while requiring the court to make use of the assistance which medical testimony may provide, is ultimately a legal one, not a medical one").

[50] Michael O'Hear, Actuarial Risk Assessment at Sentencing: Potential Consequences for Mass Incarceration and Legitimacy, 38 Beh. Sci. & L. 193, 195-196 (2020).

[51] Brandon L. Garrett & Megan Stevenson, Open Risk Assessment, 38 Beh. Sci. & L. 279 (2020).

[52] 509 U.S. 579 (1993).

[53] See, e.g., Tex. Code Crim. Proc., Art. 37.071(b)(2) (defining dangerousness aggravating circumstance as "a probability that the individual will commit criminal acts of violence that constitute a continuing threat to society."); Long v. State, 2009 WL 960598, at *3 (Tex. Crim. App. 2009) (refusing to quantify or define probability).

[54] Nicolas Scurich, The Case Against Categorical Risk Assessments, 37 Beh. Sci. & L. 554, 556-557 (2018) (reporting studies in which the "high risk" designation among clinicians ranged from 38% to 100%, produced a mean among evaluators of 63.% with a very large standard deviation of 23.2%, and a range of 5 to 100% from a sample of judges and forensic clinicians). See also Cecelia Klingele, Making Sense of Risk, 38 Beh. Sci. & L. 218, 222-223 (2020) (noting the dangers of failing to look behind categorical risk judgments).

[55] Daniel Krauss, Gabriel I. Cook &. Lucas Klapatch, Risk Assessment Communication Difficulties: An Empirical Examination of the Effects of Categorical versus Probabilistic Risk Communication in Sexually Violent Predator Decisions, 36 Behav. Sci. & L. 532, 544 (2018).

[56] See supra text accompanying notes 26-28. Some studies have suggested that white people use and sell drugs at even higher rates than black people. Jonathan Rothwell, How the War on Drugs Damages Black Social Mobility, Brookings Inst. (Sept. 30, 2014), http://www.brookings.edu/ blogs/social-mobility-memos/posts/2014/09/30/how-the-war-on-drugs-damages-black-socialmobility [https://perma.cc/9KX3-UJ6B]. See also, Katherine Beckett, Kris Nyrop & Lori Pfingst, Race, Drug and Policing: Understanding Disparities in Drug Delivery Arrests, 44 Criminol. 105 (2006); Klingele, supra note 54, at 221 ("The factors that predict capture are tied more closely to the intensity of population surveillance and ease of detection by police than they are to the actual prevalence of the criminal behaviors themselves.").

[57] This differential is due to a number of reasons that are usually ambiguous as to the viability of the arrest, including insufficient evidence (suggesting innocence, but possibly due, e.g., to a reluctant witness) and deals with the prosecution (suggesting guilt but possibly due to a desire to avoid a greater sentence at trial).

[58] See Johnson v. United States, 135 S.Ct. 2551 (2015) (noting the large number of crimes that could be called violent, in the course of finding unconstitutional language in the Armed Career Criminal Act that authorized an additional 15 year sentence for *"*conduct that presents a serious potential risk of physical injury to another*").*

[59] Andrew Ashworth & Lucia Zedner, Preventive Justice 260 (2014) (arguing that sentence enhancements based on risk should not occur unless the person "is adjudge to present a very serious danger to others" and the person "has a previous conviction for a very serious offence"). The revisions to the Model Penal Code Sentencing Provisions are in accord. See Model Penal Code: Sentencing § 1.02(2), Comment b (2020) ("In the Code, the term is meant to include offenses that cause or risk serious injuries to crime victims, and to rule out less serious crimes, but its precise boundaries are left to the legislatures, sentencing commissions, and courts of each state.").

[60] Rhys Hester, Risk Assessment Savvy: The Imperative of Appreciating Accuracy and Outcome, 38 Beh. Sci. & L. 246, 250-252 (2020) (demonstrating statistically how changing the outcome measure from general recidivism to violent recidivism can drastically reduce the probability of reoffending, e.g., from 79% to 18% in the highest risk category).

[61] Richard B.A. Coupland & Mark E. Olver, Assessing Protective Factors in Treated Violent Offenders: Associations with Recidivism Reduction and Positive Community Outcomes, 20 Psychol. Assessment 493 (2020).

[62] See, e.g., Greg Pogarsky & Alex R. Piquero, Can Punishment Encourage Offending? Investigating the "Resetting" Effect, 40 Journal of Research in Crime and Delinquency 95 (2003); Travis C. Pratt & Francis T. Cullen, F. T., Assessing Macro-level Predictors and Theories of Crime: A Meta-analysis, in 32 Crime and Justice: A Review of Research 373 (Michael Tonry, ed., 2005); Paula P. Smith, Claire Goggin & Paul Gendreau, The Effects of Prison Sentences and Intermediate Sanctions on Recidivism: General Effects and Individual Differences (2002); Patrice Villettaz, Martin Killias & Isabel Zoder, The Effects of Custodial vs. Non-custodial Sentences on Re-offending: A Systematic Review of the State of the Knowledge (2006).

[63] Bonta & Andrews, supra note 4, at 12 ("effectiveness is maximized when the treatment is in a community setting"); Mark Lipsey & Francis Cullen, The Effectiveness of Correctional Rehabilitation: A Review of Systematic Reviews, 3 Ann. Rev. L. & Soc. Sci. 297, 302 (2008); James Bonta, Suzanne Wallace-Capretta & Jennifer, A Quasi-experimental Evaluation of an Intensive Rehabilitation Supervision Program, 27 Criminal Justice & Behavior 312 (2000) (finding, in an evaluation of a Canadian program, that low risk offenders who received minimal levels of treatment had a recidivism rate of 15% and low risk offenders who received intensive levels of services had more than double the recidivism rate (32%), while high risk offenders who did not receive any intensive treatment services had a recidivism rate of 51% but those who did had almost half the recidivism rate (32%)); Phyllis Solomon, Jeffrey Draine, Stephen C. Marcus ,Predicting Incarceration of Clients of a Psychiatric Probation and Parole Services, 53 Psychiatric Services 50 (2002) (finding that intensive surveillance can lead to re-arrests for minor crimes that normally would not occasion incarceration). For a description of the program, similar programs, and relevant data, see Ctr. for

the Study and Prevention of Violence, Blueprints for Violence Prevention (2004), https://www.ncjrs.gov/pdffiles1/ojjdp/204274.pdf.

[64] Kevin R. Reitz, The Compelling Case for Low-Violence-Risk Preclusion in American Prison Policy, 38 Beh. Sci. & L. 207 (2020); Hester, supra note 60, at 253; Stephen D. Gottfredson & Michael Gottfredson, Selective Incapacitation?, 478 Annals of the American Acad. Of Pol. & Soc. Sci. 135 (1985); Hennessey D. Hayes & Michael R. Geeken, The Idea of Selective Release, 14 Just. Quar. 353, 368-69 (1997).

[65] Jay P. Singh, Predictive Validity Performance Indicators in Violent Risk Assessment, 31 Behav. Sci. & L. 8, 18 (2013).

[66] Id.

[67] Melissa Hamilton, Judicial Gatekeeping on Scientific Validity with Risk Assessment Tools, 38 Behav. Sci. & L. 234 (2020).

[68] Id. at 234.

[69] Id.

[70] There is no commonly used term for this concept, but category base rate is an accurate descriptor, and probably preferable to category positive predictive value, because PPV has an accepted meaning within the empirical.

[71] Kern & Farrar-Owens, supra note 6, at 165.

[72] Lin et al., supra note 19, at 4.

[73] Singh, supra note 65, at 15.

[74] See Taanvi Ramesh et al., Use of Risk Assessment Instruments to Predict Violence in Forensic Psychiatric Hospitals: A Systematic Review and Meta-analysis, 52 Eur. Psychiat. 47, 50 tbl. 3 (2018) (reporting AUCs for 9 instruments ranging from .71 to .85 for predictions of imminent offending risk and between .62 and .75 for longer-range predictions).

[75] L. Maaike Helmus & Kelly M. Babchishin, Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy, 44 Crim. Just. & Behav. 8, 12 (2017). These same authors state that other researchers are more likely to say that AUCs between .60 and .69 are poor, .70 to .79 are fair, .80 to .89 are good, and over .90 are excellent. Id. at 11. See also Sarah L. Desmarais & Jay P. Singh, Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States, Council of State Governments (2013) (analysis comparing AUC values to other measures of predictive validity (Cohen's d and odds ratios) and labeling AUCs of .55 to .63 as "fair," .64 to .71 as "good," and .71 to 1.00 as "excellent.").

[76] See, e.g., Fed. R. Evid. 401.

[77] See Pamela M. Casey et al., Nat'l Ctr. for State Courts, Using Offender Risk and Needs Assessment Information at Sentencing: Guidance for Courts from a National Working Group 29-30 (2011) ("After identifying the most promising tool for use in a jurisdiction, the supervising agency should validate the instrument on a sample that is representative of the local population before undertaking full-scale implementation").

[78] See Astrid Rossegger, Replicating the Violence Risk Appraisal Guide: A Total Forensic Cohort Study, 9 Plos One e91845 at 2 (2014) (stating that "[s]tudies that *have* investigated the goodness-of-fit between the rates of violent recidivism published in the VRAG manual and rates observed during the research have produced inconsistent findings" but finding acceptable predictive validity on European samples).

[79] See, e.g., Joel K. Cartwright et al., Predictive Validity of HCR-2), START, and STATIC-99R Assessments in Predicting Institutional Aggression among Sexual Offenders, 42 Law & Hum. Beh. 13

[80] See John Logan Koepke & David G. Robinson, Danger Ahead: Risk Assessment and the Future of Bail Reform, 93 Wash. L. Rev. 1725, 1756 (2017) ("Using one jurisdiction's data to predict outcomes in another is an inherently hazardous exercise, a challenge that is highlighted in the existing literature").

[81] Jennifer L. Skeem & Christopher Lowenkamp, Risk, Race, and Recidivism: Predictive Bias and Disparate Impact, 54 Criminol. 680, 700 (2016) ("these results indicate that risk assessment is not 'race assessment'"); Desmarais, Johnson & Jay Singh, *supra* note 21, at 216 (2016) (in metareview of 19 RAIs, of the three that permitted comparison of black and white AUCs, the measures were "identical" or "highly similar"); Jay Singh & Seena Fazel, Forensic Risk Assesment: A Metareview, 37 Crim. Just. & Beh. 965, 978 (2010) (finding that, of six meta-reviews, five found "no evidence that predictive validity varied by the ethnicity of participants").

[82] Hamilton, supra note 67, at 231-232.

[83] Id. at 236.

[84] Ewert v. Canada [2018] 2 S.C.R. 165 (Can.).

[85] See Casey et al., supra note 77, at 30 (validation "should include empirical efforts to norm the tool on different groups of offenders in the target population to ensure that the tool produces accurate risk classifications across subgroups"); Min Yang et al., The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools, 136 Psychol. Bull. 740, 741 (2010) ("[The] predictive efficacies of all tools must be eventually subjected to repeated empirical validation with client groups that differ in demographic characteristics (e.g., age, gender, socioeconomic status, ethnicity), level and type of past violence (e.g., criminal histories, sexual vs. nonsexual offenders), psychiatric diagnosis (e.g., presence of personality disorder, psychosis), intervention received (e.g., treated vs. untreated), the specific criterion being predicted (e.g., violent vs. nonviolent behavior or different types of violent behavior), environmental setting (e.g., clients residing in institutions vs. the community), countries of origin of the research, and so forth.").

[86] State v. Gordon, 919 N.W.2d 635, *9 (Iowa Ct. App.), vacated on other grounds, 921 N.W.2d 19 (Iowa S. Ct. 2018) ("Nothing in our record indicates the existence of validation studies for these tests or any cross validation for an Iowa population of offenders"); State v. Loomis, 881 N.W.2d 749, 764 (Wisc. 2016) (judges should be aware that a given instrument may not have been normed on local populations and "must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations"); Ewert v. Canada, 2 S.C.R. 165 ¶65 (Can. 2018) ("the clear danger posed by the CSC's continued use of assessment tools that may overestimate the risk posed by Indigenous inmates is that it could unjustifiably contribute to disparities in correctional outcomes in areas in which Indigenous offenders are already disadvantaged.").

[87] See Grant Duwe, Public Safety Clearinghouse, Why Inter-Rater Reliability Matters for Recidivism Risk Assessment 2 (2017), https://psrac.bja.ojp.gov/ojpasset/Documents/PB-Interrater-Reliability.pdf (explaining that few studies have adequately studied inter-rater reliability but of the few studies that exist several found "relatively low IRR [inter-rater reliability] for recidivism risk assessment tools."); Anthony W. Flores et al., Predicting Outcome with the Level of Services Inventory-Revised: The Importance of Implementation Integrity, 34 J. Crim. Justice 523, 528 (2006) (predictive validity of results under the LSI-R (LSCMI) suffered when the instrument was administered by untrained personnel).

[88] Desmarais & Singh, supra note 21, at 213.

[89] See Council for State Governments Justice Center, Practical Guidelines for the Use of Post-Conviction Risk and Needs Assessment Tools (2020) (hereafter, Practical Guidelines).

[90] Koepke & Robinson, supra note 80, at 1764-65 (pointing out that reforms designed to remind arrestees of their trial dates decreased failures-to-appear but that the effects of this innovation are not taken into account in pretrial RAIs).

[91] See Virginia Criminal Sentencing Commission, 2012 Annual Report 31, 48 (2012), http://www.vcsc.virginia.gov/2012VCSCAnnualReport.pdf (detailing the history of the first re-validation study and implementing the results of a re-validation study started in 2010); Kenneth Rose, Virginia Department of Criminal Justice Services, Risk Assessment Factsheet: Virginia Pretrial Risk Assessment Instrument 1 (2019), https://www-cdn.law.stanford.edu/wp-content/uploads/2019/06/VPRAI-Factsheet-FINAL-6-20.pdf ("Though the VPRAI was first developed in 2003, the tool was revised in 2007 to remove the 'Outstanding Warrants' factor and was further revised in 2016").

[92] See Brandon Garrett & John Monahan, Judging Risk, 108 Calif. L. Rev. 439, 487 (2020) (calling for "(1) publicly specifying the criteria for risk assessment instruments; (2) defining the relevant risks and needs to be measured; (3) making the risk instrument public and accessible to researchers; (4) presenting risk information in a comprehensible way to decision-makers; (5) structuring decision-making to make better use of that information; and (6) accompanying these reforms with ongoing monitoring, through judicial review and by making data accessible to researchers").

[93] See Practical Guidelines, supra note 89 (recommending case audits twice a year that examine fidelity to coding and scoring guidelines and the match between assessment results and case planning).

[94] See, e.g., Brandon Garrett, Alexander Jakubow & John Monahan, A Report of the Virginia Criminal Justice Policy Reform Project (2018), available atvcsc.virginia.gov/2018meetings/UVA%20Law%20School% 20-%20NVRA%20Sentencing% 20Analysis%20and%20 Judicial% 20Survey%20(Mar%201%202018).pdf.

[95] Fed.R.Evid. 702.

[96] See Practical Guidelines, supra note 89 (recommending annual "booster training" that reviews coding and scoring guidelines and involves completion of "practice cases" that produce high agreement between raters).

[97] Matthew DeMichele, et al., The Intuitive Override Model: Nudging Judges Toward Pretrial Risk Assessment Instruments 18 (2018), https://craftmediabucket.s3.amazonaws.com/uploads/PDFs/5-Intuitive-Override-Model.pdf.

[98] Id.

[99] See, e.g., State v. Loomis, 881 N.W.2d 749, 769 (Wis. 2016); Malenchik v. State, 928 N.E.2d 564, 573 (Ind. 2010).

[100] See sources cited supra notes 19-24.

[101] Jean-Pierre Guay & Genevieve Parent, Broken Legs, Clinical Overrides, and Recidivism Risk: An Analysis of Decisions to Adjust Risk Levels With the LS/CMI, 45 Crim.l Just. & Behavior 82, 97 (2018).

[102] Fred Schmidt, Sarah M. Sinclair & Solveig M. Thomasdót, Predictive Validity of the Youth Level of Service/Case Management Inventory with Youth Who have Committed Sexual and Sexual Offenses: The Utility of Professional Override, 43 Criminal Justice & Behav. 413, 413 (2016).

[103] Risk Assessment Overrides: Shuffling the Risk Deck without Any Improvements in Prediction, Crim. Just. & Behavior (forthcoming, 2021).

[104] R. Karl Hanson, What Do We Know about Sex Offender Risk Assessment, 4 Psychol., Pub. Pol'y & L. 50, 65 (1998) (recommending that clinicians be "exceedingly cautious" in making adjustments, but noting that "[t]hose skeptical of actuarial predictions will always find reasons to adjust actuarial estimates").

[105] Katherine E. McCallum et al., The Influence of Risk Assessment Instrument Scores on Evaluators' Risk Opinions and Sexual Offender Containment Recommendations, 44 Crim. Just. & Behav. 1213, 1214 (2017).

[106] See Risk Assessment Overrides, supra note 103 (finding this to be the case in a study of probation officers). But a concern about lack of resources *in the community* could also have the opposite effect. See Brandon Garrett & John Monahan, Assessing Risk: The Use of Risk Assessment in Sentencing, 103 Judicature (2019), https://judicature.duke.edu/articles/assessing-

risk-the-use-of-risk-assessment-in-sentencing/#_ednref21 ("Our findings confirm the 'treatment resource hypothesis' to be one factor accounting for the wide variation among courts and individual judges in the extent to which drug and property offenders assessed as low risk of recidivism actually receive a sentence of community treatment that does not include incarceration in a local jail.").

[107] Anne Metz et al., Valid or Voodoo: A Qualitative Study of Attorney Attitudes Towards Risk Assessment in Sentencing and Plea Bargaining (2020), https://ssrn.com/abstract=3552018.

[108] State v. Loomis, 881 N.W.2d 749, 758 (Wisc. 2016) (favorably contrasting "evidence-based sentencing" with "ad hoc decision making"); Malenchik v. State, 928 N.E.2d 564, 573 (Ind. 2010)("Having been determined to be statistically valid, reliable, and effective in forecasting recidivism, the assessment tool scores may, and if possible should, be considered to supplement and enhance a judge's evaluation, weighing, and application of the other sentencing evidence in the formulation of an individualized sentencing program appropriate for each defendant").

[109] Angèle Christin et al., Courts and Predictive Algorithms 7, Data Society (Oct. 27, 2015), http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf; Steven J. Wormith et al., The Predictive Validity of a General Risk/Needs Assessment Inventory on Sexual Offender Recidivism and an Exploration of the Professional Override, 39 Crim. Just. & Behav. 1511, 1516 (2012).

[110] Anne Metz et al., Risk and Resources: A Qualitative Perspective on Low-Level Sentencing in Virginia 15 (2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3437750 (finding that 83% of judges considered the results of the NVRA as a "validating data point" but not "dispositive," because they also consider "the facts of the case and the defendant's criminal history").

[111] Kimberly Jenkins Robinson, The Constitutional Future of Race-Neutral Efforts to Achieve Diversity and Avoid Racial Isolation in Elementary and Secondary Schools, 50 B.C. L. Rev. 277, 315 (2009) ("The Court's current approach to equal protection, which has been labeled an antidiscrimination, anticlassification, or color-blind approach, emphasizes the impropriety of government use of racial classifications."). See, e.g., Parents Involved in Cmty. Schools v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 748 (2007) (plurality opinion) ("[t]he way to stop discrimination on the basis of race is to stop discriminating on the basis of race.").

[112] Buck v. Davis, 137 S.Ct. 759, 775, 778 (2017) (holding that "it would be patently unconstitutional for a state to argue that a defendant is liable to be a future danger because of his race," and nothing that using the status of being black as a risk factor "appealed to a powerful racial stereotype" and "coincided precisely with a particularly noxious strain of racial prejudice").

[113] 429 U.S. 190 (1976).

[114] 881 N.W.2d 749 (Wis. 2016).

[115] Id. at 766. Other courts have recognized this point. See also Karsjens v. Jesson, 6 F. Supp.3d 958, 967-68 (D. Minn. 2015) (noting, in case challenging female's sex offender civil commitment programming, experts' testimony that actuarial risk tools normed on male sex offenders are inapplicable to females); In re Risk Level Determination of S.S., 726 N.W.2d 121, 123 (Minn. Ct. App. 2007) (noting expert declined to score a sexual recidivism risk tool for a female defendant as it had not been validated on women).

[116] Starr, supra note 27, at 830-836.

[117] 461 U.S. 660 (1983).

[118] Id. at 671-672.

[119] See, e.g., United States v. Flowers, 946 F. Supp. 2d 1295, 1300 (M.D. Ala. 2013) ("[R]elative wealth and poverty will inevitably have some effect on the administration of justice."); State v. Johnson, 315 P.3d 1090, 1099 (Wash. 2014) (interpreting *Bearden* to mean that a person cannot be imprisoned for failure to pay a fine).

[120] 461 U.S. at 670.

[121] Id. at 671.

[122] Personal communication with Meredith Farrar-Owens, March 25, 2020 (memo indicating that these two factors were dropped on July 1, 2013, "based on a study of new felony cases").

[123] Michael Coenen, Spillover Across Remedies, 98 Minn. L. Rev. 1211, 1229-1230 (2014) (stating that "the discriminatory purpose rule applies with equal force in criminal cases," noting, for instance, that notwithstanding powerful evidence of discrimination against blacks in cocaine prosecutions, "Washington v. Davis and its progeny stood as an impenetrable barrier to equal protection relief, as black crack-cocaine defendants never managed to gather enough evidence to satisfy the discriminatory purpose requirement.").

[124] Cary Coglianese & David Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era, 105 Geo. L.J. 1147, 1193 (2017).

[125] Ohio Rev. Code Ann. 2929.11(C) ("A court that imposes a sentence upon an offender for a felony shall not base the sentence upon the race, ethnic background, gender, or religion of the offender").

[126] Tenn. Code Ann. 40-35-102(4) ("Sentencing should exclude all considerations respecting race, gender, creed, religion, national origin, and social status of the individual").

[127] State v. Loomis, 881 N.W.2d 759, 761 (Wis. 2016) (holding that, because Loomis had access to the questions the COMPAS asked and the risk assessment itself, he did not need access to "how the risk scores were determined or how the factors are weighed").

[128] Megan T. Stevenson & Christopher Slobogin, Algorithmic Risk Assessments and the Double-Edged Sword of Youth, 96 Wash. U. L. Rev. 681, 688-698 (2018).

[129] 18 U.S.C. § 3631(b)(4) (requiring the Attorney General, inter alia, to "on an annual basis, review, validate, and release publicly on the Department of Justice website the risk and needs assessment system").

[130] Department of Justice, The First Step Act Risk and Needs Assessment System—Update 14 (January, 2020) (hereafter First Step Act—Update), available at https://www.bop.gov/inmates/fsa/docs/the-first-step-act-of-2018-risk-and-needs-assessment-system-updated.pdf

[131] Nicholas Diakopoulos, We Need to Know the Algorithms the Government Uses to Make Important Decisions About Us, *Conversation* (May 23, 2016, 8:48 PM), https://theconversation.com/we-need-to-know-the-algorithms-the-government-uses-to-make-important-decisions-about-us-57869 (noting that only one state responded fully to Freedom of Information Act requests for algorithm documents and source codes).

[132] See generally, Doaa Abu Elyounes, Bail or Jail? Judicial versus Algorithmic Decision-Making in the Pretrial System, Colum. Sci. & Tech. L. Rev. (forthcoming, 2020) (describing various types of machine learning algorithms and possible difficulties with discerning how they work, but also noting that most RAIs today rely on "traditional regression analysis," id. at 7, and are transparent with respect to the factors considered and the weight they are given, id. at 76-77).

[133] Andrew Selbst & Simon Barocas, The Intuitive Appeal of Explainable Machines, 87 Ford. L. Rev. 1085, 1091 (2018).

[134] Alexandra Chouldechova, Estella Loomis McCandless & Kristian Lum, The Present and Future of AI in Pretrial Risk Assessment Instruments (June, 2020) ("AI technologies are not likely to achieve considerably greater predictive accuracy than currently available risk assessment instruments."); Cynthia Rudin and Berk Ustun, Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice, 48 Interfaces 449 (2018) (absent the inclusion of more complex unstructured input data, simple models that rely on a small set of factors such as age and prior system involvement can perform just as well); Elaine Angelino et al., Learning Certifiably Optimal Rule Lists for Categorical Data, 18 J. Machine Learning Research 8753 (2017).

[135] Cynthia Rudin, Caroline Wang & Beau Coker, The Age of Secrecy and Unfairness in Recidivism Prediction, 2 Harvard Data Science Rev.1 (2020), https://doi.org/10.1162/99608f92.6ed64b30 (pointing out the many ways in which opaque RAIs like COMPAS complicate evaluations of fairness and accuracy). Selbst & Barocas point out that, with respects to providing transparency, "researchers have developed at least three different ways to respond to the demand for explanations: (1) purposefully orchestrating the machine learning process such that the resulting model is interpretable; (2) applying special techniques after model creation to approximate the model in a more readily intelligible form or identify features that are most salient for specific decisions; and (3) providing tools that allow people to interact with the model and get a sense of its operation." Selbst & Barocas, supra note 133, at 1110.

[136] David G. Robinson & Logan Koepke, Civil Rights and Pretrial Risk Assessment Instruments 11 (2019). See generally Kleinberg et al., supra note 37 (arguing that "all the components of an algorithm (including the training data) must be stored and made available for examination and experimentation").

[137] 430 U.S. 360 (1977).

[138] Id. at 361.

[139] 353 U.S. 53 (1961).

[140] See Z. A. G. Perez, Piercing the Veil of Informant Confidentiality: The Role of In Camera Hearings in the *Roviaro* Determination, 46 Am. Crim. Law Rev. 179, 202–13 (2009) (describing Federal Circuit of Appeals approaches to *Roviaro*).

[141] Id.

[142] Danielle Keats Citron, Technological Due Process, 85 Wash. U. L. Rev. 1249, 1290-91 (2008) ("the public [and] government actors are to influence policy when it is shrouded in closed code.").

[143] See Rebecca Wexler, Life, Liberty and Trade Secrets: Intellectual Property in the Criminal Justice System, 17 Stan. L. Rev. 1343, 1403-13 (2018) (arguing that, given protective orders and other procedural devices, "trade secret protection is unnecessary in criminal cases"); see also, Danielle Keats Citron & Frank Pasquale, The Scored Society: Due Process for Automated Predictions, 89 Wash. L. Rev. 1, 26 (2014) ("There is little evidence that the inability to keep such systems secret would diminish innovation." It is noteworthy that intellectual property claims have also been rejected in civil cases when the potential for error is "obvious" and "substantial." See K.W. v. Armstrong, 180 F.Supp3d. 703, 716-17 (D. Idaho 2016).

[144] See generally, Margot E. Kaminski, The Right to Explanation, Explained, 34 Berkeley Tech. L.J. 189, 209-217 (2019) (reviewing the literature).

[145] Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, 2016 O.J. (L 119) 1 (EU), articles 13-15.

[146] State v. Guise, 921 N.W.2d 26 (Iowa 2018) (Appel, J., concurring).

[147] Nancy King, Procedure at Sentencing, in Oxford Handbook of Sentencing and Corrections 317, 320 (2012).

[148] McGautha v. California, 402 U.S. 183, 218-220 (1971).

[149] Judges have discretion to deny experts even at trial. See, e.g., Moore v. State, 841 A.2d 31 (Md. Spec. App. 2004) (no violation of indigent defendant's rights where his request for funding of expert was denied, because "independent" lab tested DNA); Finn v. State, 558 S.E.2d 717 (Ga. 2002) (defendant's rights not violated by denial of funds for his own expert); People v. Leonard, 569 N.W.2d 663 (Mich. App. 1997) (same).

[150] Daniel A. Krauss & Nicholas Scurich, Risk Assessment in the Law: Legal Admissibility, Scientific Validity, and Some Disparities Between Research and Practice, 31 Behav. Sci. L. 215, 220 (2013) (19 states hold that admissibility rules governing expert evidence do not apply at sentencing or that the rules are relaxed sufficiently to allow testimony based on RAIs).

[151] Greenholtz v. Inmates of Nebraska Penal and Correctional Complex, 442 U.S. 1, 15 (1979) (no right to counsel or cross-examine at parole hearings); Meachum v. Fano, 427 U.S. 215, 225 (1976) (transfer from a low security to a high security setting can be a purely administrative decision).

[152] Richard B.A. Coupland & Mark E. Olver, Assessing Protective Factors in Treated Violent Offenders: Associations With Recidivism Reduction and Positive Community Outcomes, 32 Psychological Assessment 493 (2020).

[153] See, e.g., Ky. Rev. Stat. § 439.335(1) ("In considering the granting of parole and the terms of parole, the parole board shall use the results from an inmate's validated risk and needs assessment and any other scientific means for personality analysis that may hereafter be developed.").

[154] See, e.g., Kansas Sentencing Commission, Justice Reinvestment Initiative in Kansas (2015) at 3; Utah Sentencing Commission, 2017 Adult Sentencing and Release Guidelines (2017).

[155] Wash. Rev Code § 9.94A.500 (2013).

[156] Arizona Judicial Branch, Presentence Report, at www.azcourts.gov/apsd/Evidence-Based-Practice/Presentence Report.

[157] Pretrial Detention Reform Workgroup, Pretrial Detention Reform: Recommendations to the Chief Justice 52-53 (2017), http://www.courts.ca.gov/documents/PDRReport-20171023.pdf.

[158] 18 U.S.C. § 3632(f)(4).