Principles of Risk Assessment: Sentencing and Policing

Christopher Slobogin*

Risk assessment—measuring an individual's potential for offending—has long been an important aspect of criminal justice, especially in connection with sentencing, pretrial detention, and police decision-making. To aid in the risk assessment inquiry, a number of states have recently begun relying on statistically-derived algorithms called "risk assessment instruments" (RAIs).[1] RAIs are generally thought to be more accurate than the type of seat-of-the-pants risk assessment in which judges, parole boards, and police officers have traditionally engaged.[2] But RAIs bring with them their own set of controversies.

In recognition of these concerns, this brief paper proposes three principles—the fit principle, the validity principle, and the fairness principle—that should govern risk assessment in criminal cases. After providing examples of RAIs, it elaborates on how the principles would affect their use in sentencing and policing. While space constraints preclude an analysis of pretrial detention, the discussion should make evident how the principles would work in that setting as well.

---

* Milton Underwood Professor of Law, Vanderbilt University Law School. This paper was prepared for The Ohio State University Moritz College of Law Round Table on Big Data and Criminal Law. The author would like to thank the participants at that conference as well as participants in faculty workshops at the University of Utah S.J. Quinney College of Law and the University of Washington Law School for their input on this paper.

[1] On sentencing, see Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 566–67 (2015) (providing cites); on pretrial detention, see Megan T. Stevenson, *Assessing Risk Assessment in Action* 8–15, (George Mason Legal Research Paper No. LS 17-25, 2018), https://ssrn.com/abstract=3016088 [https://perma.cc/BDF6-Z5UY]; on policing, see Andrew Guthrie Ferguson, *Policing Predictive Policing*, 94 WASH. U. L. REV. 1109, 1126–42 (2017).

[2] N. Zoe Hilton, Grant T. Harris & Marnie E. Rice, *Sixty-Six Years of Research on the Clinical Versus Actuarial Prediction of Violence*, 34 COUNSELING PSYCHOLOGIST 400, 400–01 (2006). *But see* Julia J. Dressel, Accuracy and Racial Biases of Recidivism Prediction Instruments, (May 31, 2017) (senior honors thesis, Dartmouth Computer Science Technical Report TR2017-822), http://www.cs.dartmouth.edu/farid/downloads/publications/jdthesis17.pdf [http://perma.cc/TQM2-7E43] (finding no difference in lay and statistical predictive accuracy, but using a methodology that in essence converted the lay prediction into an algorithm).

I. RISK ASSESSMENT INSTRUMENTS

Today, there are a huge number of RAIs, some developed by the government and some by researchers at universities or private companies.[3] In the space allotted, justice cannot be done to the wide array of such instruments. But the three RAIs described here provide enough background to acquaint the reader with the nature of data-based risk assessment and different ways of approaching it.

A relatively new RAI is the Oxford Risk of Recidivism Tool, nicknamed OxRec.[4] According to its initial validation study, the OxRec was able to identify a "high risk" group, 60% of whom committed a violent offense within a two-year period, a "medium risk" group with a 30% recidivism rate over two years, and a "low risk" group that had less than a 10% recidivism rate within that period.[5] In the world of risk assessment, these are good results. An instrument that can reliably differentiate between groups with 60% and 10% recidivism rates—and whose high risk group includes only 40% non-recidivists and whose low risk group includes only 10% recidivists—is state of the art.

The OxRec relies on weighted "risk factors," which is typical of RAIs. But the OxRec is noteworthy because it considers so *many* risk factors, including environmental variables that other instruments do not consider. Here is the full list of OxRec risk factors: male sex; unemployed before prison; young age; non-immigrant status; previous prison sentence of short duration; violent index crime; previous violent crime; never married; fewer years of education; low disposable income; alcohol use disorder; drug use disorder; any mental disorder; any severe mental disorder; and "high neighbourhood deprivation," which is determined using rates or measures of welfare recipiency, migration status, divorce, educational levels, residential mobility, crime, and disposable income within the individual's neighborhood.[6] The rationale for inclusion of these factors, some of them counter-intuitive, is strictly statistical.

A much older instrument, the Violence Risk Appraisal Guide (VRAG), is used extensively in Canada and in several U.S. jurisdictions.[7] It relies on twelve risk factors, having to do with the individual's score on the Psychopathy Checklist (a measure of psychopathy that takes into account criminal history); elementary school misconduct; diagnosis (with personality disorders positively, and

---

[3] T. Douglas et al., *Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data*, 42 EUR. PSYCHIATRY 134, 134 (2017), http://dx.doi.org/10.1016/j.eurpsy.2016.12.009 [http://perma.cc/HZZ6-DWAT] (stating there are over 200 such tools).

[4] Seena Fazel et al., *Prediction of Violent Reoffending on Release from Prison: Derivation and External Validation of a Scalable Tool*, 3 LANCET PSYCHIATRY 535, 540 (2016) (referring to OxRec at https://oxrisk.com/oxrec/ [http://perma.cc/A25Q-T92V]).

[5] *Id.* at 541 fig.4.

[6] *Id.* at 537 tbl.1.

[7] For a description of the instrument and relevant research, see Grant T. Harris et al., *Prospective Replication of the* Violence Risk Appraisal Guide *in Predicting Violent Recidivism Among Forensic Patients*, 26 LAW & HUM. BEHAV. 377 (2002).

schizophrenia negatively, correlated with risk); age; presence of parents in home before age 16; performance on conditional release (if any); non-violent offenses; marital status; victim injury; victim gender; and history of alcohol abuse.[8] The evaluator assigns a numerical score in connection with each risk factor according to a statistically-derived table. Scores can range from less than -21 to more than 28, with the lowest score predicting a 0% chance of violent offending within seven years and the highest score predicting a 100% chance of violent offending within that period.[9] Seven other "bins" or ranges are associated with recidivism probabilities of 8% through 76%.[10]

One measure of the accuracy of RAIs like the VRAG is provided by the "receiver operating characteristic curve," which is derived by plotting the true positive rate over the false positive rate.[11] If the resulting curve follows the left vertical axis and then the upper horizontal axis (like a non-capitalized "r"), the area under the curve (AUC) would be 1.0, indicating that the instrument is 100% accurate. If, instead, the curve ends up being a 45-degree diagonal from the lower left corner to the upper right corner of the plot, the AUC would be 0.5, indicating that the RAI is no better than chance at differentiating true positives from true negatives. A typical AUC value for the VRAG is 0.75,[12] indicating that there is a 75% chance that a person who recidivates received a higher score on the VRAG than a person who did not recidivate.

The third instrument described here is the HCR-20.[13] As the name implies, this RAI consists of 20 risk factors, ten having to do with *h*istorical matters, five relating to *c*linical symptoms, and five relating to *r*isk management or treatment. The historical factors are previous violence; age at first violent incident; relationship instability; employment problems; substance use problems; major mental illness; psychopathy; early maladjustment; personality disorder; and prior supervision failure.[14] The clinical factors are lack of insight; negative attitudes; active symptoms of major mental illness; impulsivity; and unresponsiveness to treatment.[15] The risk management factors are unfeasibility of plans; exposure to destabilizers; lack of personal support; noncompliance with remediation attempts;

---

[8] *Id.* at 378.

[9] *Id.* at 385 tbl.2.

[10] *Id.*

[11] *See* Douglas Mossman, *Assessing Predictions of Violence: Being Accurate About Accuracy*, 62 J. CONSULTING & CLINICAL PSYCHOL. 783, 784–85 (1994) (describing this method of measuring the accuracy of risk assessment).

[12] *See* Majid Bani-Yaghoub et al., *A Time Series Modeling Approach in Risk Appraisal of Violent and Sexual Recidivism*, 34 LAW & HUM. BEHAV. 349, 359 (2010).

[13] For a description of this instrument and accompanying research, see Kevin S. Douglas & Christopher D. Webster, *The HCR-20 Violence Risk Assessment Scheme: Concurrent Validity in a Sample of Incarcerated Offenders*, 26 CRIM. JUST. & BEHAV. 3, 8 (1999).

[14] *Id.*

[15] *Id.*

and stress.[16] Each of the 20 factors is scored on a scale of 0–2, so that the maximum total score is 40 (although the developers of the HCR-20 counsel that a strictly mathematical assessment should be avoided and that, instead, individuals should simply be characterized as "high," "medium," or "low" risk).[17]

One key difference between the HCR–20 and the other two instruments is that the HCR-20 explicitly looks at the individual's potential for rehabilitation and likelihood of following treatment plans. Research on the HCR-20 indicates that it has AUC values similar to or higher than the VRAG.[18] One study found that individuals with scores of 1–14 on the HCR-20 reoffended at about an 11% rate within two years, while those with scores of 27 or higher reoffended at about a 75% rate within two years.[19]

These three instruments are relatively typical. Some RAIs rely on fewer risk factors, and all vary in the extent to which they consider "static" factors that the defendant can do nothing about (like prior crimes or age) and "dynamic" risk factors that are changeable (like substance abuse or impulsivity).[20] But the foregoing account of RAIs is sufficient for purposes of assessing their role in sentencing.

II. THREE PRINCIPLES OF RISK ASSESSMENT

The question this section addresses is whether instruments like the OxRec, the VRAG, and the HCR-20 should play a role in deciding whether a person should be incarcerated, receive a sentence enhancement, or be eligible for early release (RAIs and policing are taken up in the next section). One could easily answer this question negatively on the ground that risk should *never* play a role in sentencing, for a number of philosophical and practical reasons that I have addressed elsewhere.[21] And even if risk is a legitimate sentencing issue, one might resist using RAIs because their quantified, mechanistic nature dehumanizes the process.

This article will assume, however, that risk is permissibly considered at the dispositional stage and that, because they produce more accurate, consistent, and transparent conclusions about risk, RAIs should be preferred over unstructured clinical judgment. Even on these assumptions, the usefulness of RAIs must be

---

[16] *Id.*

[17] *Id.*

[18] Laura S. Guy et al., *Influence of the HCR-20, LS/CMI, and PCL-R on Decisions About Parole Suitability Among Lifers*, 39 LAW & HUM. BEHAV. 232, 235–38 (2015).

[19] Kevin S. Douglas et al., *Assessing Risk for Violence Among Psychiatric Patients: The HCR-20 Violence Risk Assessment Scheme and the Psychopathy Checklist: Screening Version*, 67 J. CONSULTING & CLINICAL PSYCHOL. 917, 924–25, tbl.7 (1999).

[20] A popular RAI that consists of only six factors, all of them static and most linked to offending, is the Static-99 (Revised). *See Static-99/Static-99R*, STATIC 99 CLEARINGHOUSE, http://www.static99.org/ [http://perma.cc/C4FB-VNKC] (last visited Feb. 18, 2018) .

[21] *See* Christopher Slobogin, *Prevention as the Primary Goal of Sentencing: The Modern Case for Indeterminate Dispositions in Criminal Cases*, 48 SAN DIEGO L. REV. 1127 (2011).

carefully analyzed. The assertion here is that, in choosing RAIs, courts and parole boards should be governed by three principles—the fit principle, the validity principle, and the fairness principle. Application of these principles suggests that RAIs will rarely meet the requirements a legal system should demand from risk assessment tools.

A. *The Fit Principle*

The fit principle, which can be gleaned from the Supreme Court's opinion in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[22] posits that RAIs ought to address the precise legal question at issue. While psychologists have done good work devising RAIs, they have not always paid attention to what the law needs. Of course, the bulk of the blame for this failing rests not on the scientists but on the courts, because judges have not been clear about what the legal issues are, nor have they forced psychologists, through evidentiary rulings, to address them.

What types of questions should RAIs be answering? First, the outcome variable used in validating the RAI ought to reflect the seriousness of what is at stake—in this case, incarceration or an enhanced sentence. Thus, I would assert (along with many others)[23] that even a high risk of minor criminal behavior should not affect the decision about incarceration. If one accepts that position, research on these instruments ought to use serious violence, not just any violence, as the outcome measure. Many RAIs may fall short in this regard. For instance, the outcome measure in the original validation research for the VRAG included a simple assault.[24] A risk of that type of violence is an insufficient basis by itself to justify incarceration or sentence enhancement.

Second, the prediction period associated with the RAI should fit legal requirements. Recall that the VRAG predicts violence within a seven-year period, while the OxRec and HCR-20 predict within two years. Since many statutory sentence ranges end well before seven years, that period is far too long for any rational sentencing regime. And even two years may be too long. While a regime that delays parole hearings for several years is probably not unconstitutional,[25] parole decisions are normally based on a mixture of retributive and utilitarian considerations. In the pure preventive detention context, in contrast, the Supreme Court has held that routine periodic review is constitutionally required.[26] To the extent a sentence is preventive in nature, the review should also be routine, perhaps

---

[22] 509 U.S. 579, 591 (1993).

[23] *See, e.g.*, ANDREW ASHWORTH & LUCIA ZEDNER, PREVENTIVE JUSTICE 260 (2014).

[24] Harris et al., *supra* note 7, at 383.

[25] Garner v. Jones, 529 U.S. 244, 251 (2000) (upholding an eight-year delay against an ex post facto challenge but also noting such delay would not be permissible if it "creates a significant risk of prolonging [the offender's] incarceration").

[26] Kansas v. Hendricks, 521 U.S. 346, 363–64 (1997).

on an annual basis.[27] That would mean that neither the VRAG *nor* the OxRec or HCR-20 provide sufficient legal fit in this regard.

Third, to the extent a sentence is based on risk, a number of Supreme Court opinions can be read to require that the disposition must be the least restrictive means of achieving the state's preventive aim.[28] Algorithmic studies should help the court assess not only risk levels but also whether something less restrictive than prison, such as a halfway house, ankle bracelets, surveillance, or outpatient treatment, can achieve the state's preventive goal. In other words, risk instruments ought to address risk management as well as risk assessment. While the HCR-20 provides such information, neither the OxRec nor the VRAG do so.

Finally, RAIs ought to be able to provide specific probability estimates of an offender's risk. Ideally, groups with a very high probability of offending could be identified. Given the state of the predictive art, however, identifying groups associated with anything over a 75% chance of recidivating is probably impossible, and even if the goal is merely meeting the preponderance standard (51%), the designated group is likely to be very small in number. Some American jurisdictions have dealt with this problem through manipulating the definition of dangerousness. For instance, under the Texas death penalty scheme, the aggravating factor of dangerousness is proven only if the state can show beyond a reasonable doubt "whether there is a probability that" a capital murder offender "would commit criminal acts of violence that would constitute a continuing threat to society."[29] Technically, that language means that the state need only show beyond a reasonable doubt a 51% likelihood that the person will reoffend, a much easier task than proving beyond doubt that the person *will* reoffend. Whether such a showing suffices as a normative matter is a tough question, although Texas (and the U.S. Supreme Court)[30] has answered it in the affirmative.

B. *The Validity Principle*

The second principle that should govern algorithmic sentencing requires that risk assessment provide reliable risk estimates. Of course, under *Daubert*, at a minimum a validity requirement mandates that the instrument be developed in a methodologically sophisticated way and that its psychometric properties be

---

[27] *See id.* at 364 (citing KAN. STAT. ANN. § 59-29a08 (1994), which required annual evaluations, reports, and hearings).

[28] *See* Slobogin, *supra* note 21, at 1138–40 (describing Jackson v. Indiana, 406 U.S. 715 (1972); Youngberg v. Romeo, 457 U.S. 307 (1982); and Seling v. Young, 531 U.S. 250 (2001), and arguing that they announce a less-drastic-means requirement where the government's goal is prevention).

[29] TEX. CODE CRIM. PROC. art. 37.071.2(b)–(c) (2017).

[30] The Court upheld the Texas death penalty statute in Jurek v. Texas, 428 U.S. 262, 274–76 (1976).

evaluated on a routine basis.[31] But it should also require experts to offer, and courts to consider, more specific external and internal validity metrics.

With respect to external validity, the RAI should be normed on a population that matches the target of the intervention. The VRAG was originally normed in Canada, which made its use problematic in the U.S. until it was validated on more diverse U.S. populations.[32] Likewise, an instrument normed on sex offenders should not be used to predict recidivism among other types of offenders. Ideally, the RAI's reference group will be highly similar to the individual being assessed in terms of both demographic characteristics and criminal charges.[33]

Internal validity is equally important. On this score, courts could, and I would argue should, demand AUC values of somewhere near 0.75 when the instrument is being used to adjust confinement. That number is not chosen arbitrarily. In *Addington v. Texas*,[34] the Supreme Court held that a mentally ill person may not be detained on dangerousness grounds on less than clear and convincing evidence, which is conventionally quantified as a 75% level of certainty. If a person cannot be involuntarily *hospitalized* without that degree of confidence, at least that much should be required before an offender may be preventively detained *in prison*. In AUC terms, that means that courts should not only require whatever level of probability the clear and convincing standard dictates under the fit principle but also require that any RAI they use accurately distinguishes high and low risk offenders roughly 75% of the time.

C. *The Fairness Principle*

Fairness is, of course, a broad concept, and could include the fitness and validity principles just discussed. Here, however, it is meant to focus solely on concerns triggered by the traditional assumption that criminal justice dispositions should be related to blameworthy conduct. For instance, in *Buck v. Davis*,[35] the Supreme Court stated:

> It would be patently unconstitutional for a state to argue that a defendant is liable to be a future danger because of his race. . . . [That would be] a disturbing departure from a basic premise of our criminal justice system: *Our law punishes people for what they do, not who they are*.[36]

---

[31] *See* Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 592–94 (1993) (requiring the basis of scientific testimony be subject to scientific testing that provides error rates).

[32] *See* Harris et al., *supra* note 7, at 381.

[33] The argument has been made that, even under such circumstances, group-to-individual assessments are not possible. But that argument is fallacious. *See* Peter B. Imrey & A. Philip Dawid, *A Commentary on Statistical Assessment of Violence Recidivism Risk*, 2 STAT. & PUB. POL'Y 25 (2015).

[34] 441 U.S. 418, 431–33 (1979).

[35] 137 S. Ct. 759 (2017).

[36] *Id.* at 775, 778 (emphasis added).

Taken literally, the italicized language would seem to prohibit all sentences based on risk—which is ultimately about status, not conduct—or at least only allow such sentences if the relevant risk factors focus on prior crimes. However, the Supreme Court probably does not mean its statement to be taken literally. In other cases (and implicitly in *Buck* itself), the Court has even upheld death sentences based on dangerousness.[37] *Buck v. Davis* appears to be a case about race, not risk.

That does not mean that the concern expressed by the Court is irrelevant. The concern comes in two forms—the discrimination claim and the dignity claim. The first part of the above-quoted language in *Buck* is most clearly related to the discrimination claim and, on the surface, it seems like a strong one. Although RAIs do not explicitly consider race, they usually do distinguish between offenders based on other immutable or near-immutable characteristics, such as gender, age, diagnosis, and various factors related to poverty. So, one might argue, sentences based on these RAIs discriminate on the basis of suspect, quasi-suspect, or quasi-quasi-suspect classes.[38]

One response to this concern is that statistician- or mechanistically-derived RAIs demonstrate no "animus" toward any of these classes, a showing that is usually required before constitutional discrimination is found.[39] A second is that even intentional racial discrimination is permissible when necessary to achieve a compelling state interest and that other forms of intentional discrimination usually require only a rational or significant justification—here, protecting the public and efficiently allocating resources.[40]

But one does not have to wade into that morass to see why the discrimination claim is hard to make out. The Wisconsin Supreme Court's opinion in *State v. Loomis*[41] makes the point. In *Loomis*, the court was faced with a challenge to an RAI called the COMPAS, which, like the OxRec, includes maleness as a risk factor. To the argument that a sentence cannot be based on such a characteristic,

---

[37] *See, e.g.*, Barefoot v. Estelle, 463 U.S. 880 (1983). In *Buck*, the expert relied on seven "statistical factors"; race was the only factor the Court found constitutionally impermissible. 137 S. Ct. at 768, 775.

[38] This is the argument of Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 805 (2014) ("I show that several of the variables that many of the instruments use raise serious constitutional and normative concerns, and I review the empirical literature to show that the instruments do not advance state interests sufficiently to overcome those concerns.").

[39] Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1193 (2017) ("Equal protection challenges to machine learning will, in short, likely fail at the first step of analysis that demands a finding that algorithms that include or analyze class-related variables are intentionally discriminatory.").

[40] *See generally*, Christopher Slobogin, *Risk Assessment*, *in* THE OXFORD HANDBOOK OF SENTENCING AND CORRECTIONS 196, 204–05 (Joan Petersilia & Kevin R. Reitz eds., 2012).

[41] 881 N.W.2d 749 (Wis. 2016).

the court stated: "[I]t appears that any risk assessment tool which fails to differentiate between men and women will misclassify both genders."[42] Because the removal of gender from its calculus would mean that the COMPAS would actually lead to inaccurate gender distinctions in sentencing (with women being rated as higher risk and men as lower risk than they actually are), the defendant's claim failed.

A more compelling discrimination-related concern is that RAIs may *inaccurately* rely on immutable factors. This was the argument made in a *ProPublica* article reporting a study about the COMPAS—the same RAI that was at issue in *Loomis*—showing that the instrument produced disproportionately more false positives among blacks than whites.[43] The response to this concern is more complicated but boils down to this: if African-Americans are more likely to commit crime than whites, a well-constructed RAI that relies heavily on prior crimes will inevitably produce a greater percentage of false positives among blacks.[44] Trying to reduce those false positives will probably increase the percentage of false negatives who are black and also increase the number of false positives who are white.

So the real question for the statistician is whether the predicate condition stated above—that African Americans commit more crimes than other racial groups—is correct. If police arrest, prosecutors charge, or juries and judges convict in racially-driven ways, it may not be. Or perhaps the predicate is correct with respect to some crimes like drug offenses but not others.[45] Ultimately, this question needs to be resolved if the discrimination concern is to be taken seriously.

The second fairness concern focuses not on discrimination per se but on dignity, or as the Supreme Court put it in *Buck*, the "basic premise of our criminal justice system [that] [o]ur law punishes people for what they do, not who they

---

[42] *Id.* at 766.

[43] *See* Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [http://perma.cc/JZ2P-UJ6H].

[44] *See* Avi Feller et al., *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear.*, WASH. POST (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.c3eec3904d97 [http://perma.cc/3BRK-F3BA] (making the mathematical point that even if "[w]ithin each risk category, the proportion of defendants who reoffend is approximately the same regardless of race . . . if black defendants have a higher overall recidivism rate, then a greater share of black defendants will be classified as high risk.").

[45] Lowenkamp and Skeem's study of the Post Conviction Risk Assessment tool found no evidence of "bias predicting bias" (i.e., "biased criminal history records predicting biased future police decisions") and concluded that criminal history is not a proxy for race. Jennifer Skeem & Christopher T. Lowenkamp, *Risk, Race, & Recidivism: Predictive Bias and Disparate Impact* 29, 35 (2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2687339 [http://perma.cc/DX9D-HFWK]. But they did not separate out specific offenses within criminal history.

are."[46] As Andrew von Hirsch framed the point over thirty years earlier, "[u]nless the person actually made the wrongful choice he was predicted to make, he ought not be condemned for that choice—and hence should not suffer punishment for it."[47] As suggested above, perhaps that sentiment would permit not only punishment for the current crime but enhancement of that sentence based on prior crimes—an issue about which von Hirsch and others have spilled much ink.[48] But it would not permit punishment to be based on anything else. Thus, risk factors such as gender, diagnosis, parental presence, employment status, and marital status—even if not discriminatory—would be off-limits because of their insult to autonomy. Not only are these factors not crimes, but many cannot even be called "behavior" chosen by the individual. And the OxRec's neighborhood score moves even further from *Buck*'s "basic premise" by considering the status and behavior of others, over which the offender clearly has no control.

That argument has a strong intuitive appeal, even in a sentencing regime which, as is the case in every American jurisdiction,[49] permits risk to influence release only within a retributively-defined sentencing range. However, *if*, as this article is assuming, risk is a legitimate sentencing factor, the argument is irrelevant; the premise that punishment is only about what people have done no longer applies (as the Supreme Court's affirmation of death sentences based on dangerousness confirms). Risk assessments are orthogonal to culpability assessments, both conceptually (the first is forward-looking, the second backward-looking), and practically (for instance, a single prior robbery conviction might call for more enhancement on desert grounds than on risk grounds).

There are also two practical problems with von Hirsch's stance. First, removal of all non-crime factors from an RAI is likely to substantially reduce accuracy. And second, as noted above, it is likely to increase discrimination. A young male with psychopathic tendencies and one prior crime represents a much higher risk than an older female suffering from schizophrenia who has committed the same crime; yet, under von Hirsch's approach, both would be treated identically.

A more nuanced approach would balance the incremental validity provided by a given risk factor with fairness concerns. As *Buck v. Davis* held, race should never be a risk factor.[50] Other noncriminal risk factors should be included in an RAI only if they appreciably improve predictive validity. This limitation would probably still permit reliance on variables such as age and gender, since they appear to improve accuracy significantly. Marital and employment status, in

---

[46] Buck v. Davis, 137 S. Ct. 759, 778 (2017).

[47] ANDREW VON HIRSCH, PAST OR FUTURE CRIMES: DESERVEDNESS AND DANGEROUSNESS IN THE SENTENCING OF CRIMINALS 11 (1985).

[48] *See, e.g.*, *id.* at 131–36.

[49] Richard S. Frase, *Theories of Proportionality and Desert*, *in* THE OXFORD HANDBOOK OF SENTENCING AND CORRECTIONS, *supra* note 40, at 131, 144–46.

[50] *Buck*, 137 S. Ct. at 775.

contrast, may not add much incremental validity and thus might be impermissible considerations.[51] The same might be true of risk factors based on neighborhood. Remember, however, that the factors included in RAIs are there because they are statistically correlated with risk. Thus, a normative judgment must be made about when a level of correlation is so low it requires a factor's exclusion.

To minimize further any affront to dignity associated with RAIs, risk assessment should be based as much as possible on dynamic or "causal risk factors," such as drug abuse or impulsivity (a goal better achieved by the HCR-20 than the OxRec or the VRAG). These are risk factors that can be changed through intervention and thus focus on traits that the person can do something about. This aspect of the fairness principle dovetails with the fit principle's requirement that algorithmic risk assessment provides output relevant to risk management. Also consistent with this point, researchers should endeavor to include in their algorithm protective factors that reduce risk, as the VRAG does with schizophrenia. Further, as a procedural matter, defendants should be able to present their own evidence of protective factors.[52]

D. *Summary*

The fit, validity, and fairness principles are very demanding in their idealized form. Some leniency in their application will be necessary if RAIs are to be used at all. But courts and parole boards making evidentiary decisions about RAIs or basing decisions on them ought to ensure these principles heavily influence outcomes.

III. THE THREE PRINCIPLES AND POLICING

In recent years, RAIs have also crept into the investigative phase of the criminal justice system. The lessons learned from sentencing are directly applicable to this setting. If applied conscientiously, the fit, validity, and fairness principles would place significant restrictions on the use of algorithms in policework, just as they would at sentencing.

Two examples of policing RAIs illustrate the challenge. The Chicago Police Department's "Heat List" relies on eleven risk factors, such as criminal history, parole status, and gang status, to generate "risk scores" from 1 to 500, with 500

---

[51] For an example of a bivariate relationship between violence and several types of risk factors based on a study focused on the relationship of mental disorder and violence, see MACARTHUR RESEARCH NETWORK ON MENTAL HEALTH & THE LAW, THE MACARTHUR VIOLENCE RISK ASSESSMENT STUDY (Apr. 2001), http://macarthur.virginia.edu/risk.html [http://perma.cc/5QLN-3ETP].

[52] *See* CHRISTOPHER SLOBOGIN, PROVING THE UNPROVABLE: THE ROLE OF LAW, SCIENCE, AND SPECULATION IN ADJUDICATING CULPABILITY AND DANGEROUSNESS 125–29 (2006) (arguing the state's proof of risk should be limited to probability estimates based on RAIs unless the defendant proffers clinical information).

being the highest risk.[53] Various private companies claim to be able to do something similar, with instruments boasting names like Digital Stakeout, Predpol, HunchLab, and Beware. Beware, developed by a company called Intrado, purports to analyze billions of data points about an individual, including property records, commercial databases, recent purchases, and social media posts, to assign "threat scores" within a matter of seconds.[54]

The idea behind most of these devices is that they come into play after police identify a possible wrongdoer using traditional means, through observation of suspicious activity or eyewitness reports. The RAI is then used to help figure out whether to surveil, stop and frisk, or arrest the individual. If the algorithm is combined with facial recognition technology, the officer can discover the person's risk level without even having a name. These instruments are touted as a way of making policing safer and also less intrusive, since cops should not frisk a low-risk person (at least in theory). But the fit, validity, and fairness principles would curb their use in a number of ways.

As in the sentencing context, the fit principle would require that policing RAIs be aimed at predicting risk of *serious* criminal activity, at least at the felony level. Further, in contrast to the sentencing context, in the investigative setting the Supreme Court has generally demanded that the danger predicted is imminent,[55] meaning the algorithm should be used only to identify either an incipient crime hot spot or, as just discussed, the risk level of a person who is linked to a recent or soon-to-occur crime by virtue of being in the relevant vicinity. Otherwise, police could use RAIs to confront the same person repeatedly without any objective indicator that the confrontation is necessary at that particular point in time. In sentencing, the impact of the risk assessment is automatically limited by the requirement of a conviction. The principle of legality, if not Fourth Amendment case law, demands something similar in the investigative setting.[56]

With respect to the validity principle, the police should have to demonstrate that the RAI is validated on a relevant population and can generate hit rates (true positives) sufficient to justify the nature of the action the police plan to take. If police want to arrest based on the RAI, for instance, the algorithm should have a

---

[53] *See* Jessica Saunders, Priscillia Hunt & John S. Hollywood, *Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot*, 12 J. EXPERIMENTAL CRIMINOLOGY 347 (2016) (describing Chicago's Strategic Subjects List and reporting a study finding it was an ineffective crime-fighting technique).

[54] Justin Jouvenal, *The New Way Police Are Surveilling You: Calculating Your Threat 'Score,'* WASH. POST (Jan. 10, 2016), https://www.washingtonpost.com/local/public-safety/the-new-way-police-are-surveilling-you-calculating-your-threat-score/2016/01/10/e42bccac-8e15-11e5-baf4-bdf37355da0c_story.html?utm_term=.e31ca039afaf [http://perma.cc/43UD-7M9R].

[55] *See* United States v. Hensley, 469 U.S. 221, 228–29 (1985) (reasoning the principles of *Terry v. Ohio* generally limit stops on reasonable suspicion to situations involving "imminent or ongoing crimes" or a known "completed felony").

[56] *See* Christopher Slobogin, *A Jurisprudence of Dangerousness*, 98 NW. U. L. REV. 1, 17–26 (2003).

high hit rate. If, instead, they plan to stop and frisk an individual, perhaps the police should have to show that one out of three flagged by the profile have just perpetrated, are perpetrating, or will soon perpetrate a serious crime; that requirement is based on a survey finding that, on average, federal judges equated a 30% level of certainty with the reasonable suspicion required for a stop.[57] If, instead, the government plans to conduct covert surveillance, the hit rate should be proportionate to the intensity of surveillance. Surveillance of long duration might require at least a 50% hit rate, a number derived from *United States v. Jones*,[58] where five justices indicated that "prolonged" tracking—in *Jones*, it was 28 days—requires probable cause, often quantified at roughly a more-likely-than-not level of certainty.[59]

With respect to the fairness principle, the primary concern might be transparency. The risk factors in most sentencing RAIs are well-known, which explains why this paper was able to report them. In contrast, most of the companies that have developed policing RAIs, like Intrado, will not reveal their algorithms, citing proprietary interests.[60] If risk factors and their relevant weights are not disclosed, the extent to which they incorporate proxies for race—or race itself—cannot be known, and the incremental validity of a given risk factor cannot be subject to independent verification. Accordingly, courts should be empowered to force disclosure of the relevant codes, *in camera* if necessary, just as they can force disclosure of any confidential informants who are crucial to the defendant's case.[61] If, upon disclosure, it is discovered that the RAI unfairly uses proxies for race or inaccurately relies on other suspect characteristics, the RAI should be adjusted accordingly; if accuracy is thereby diminished, so be it.

What if the police want to use the RAI in a preemptive way, for instance, by tracking down people identified by the Chicago Heat List? Here all three principles come into play. The predicted risk should be serious and imminent (fit). If the police contemplate stopping people identified by the Heat List, the predicted hit rate should be in the 30% range; if instead, they only plan to conduct short-term surveillance, a lower hit rate would suffice (validity). But to ensure fairness, police should treat everyone identified by the RAI equally, meaning either that all

---

[57] C.M.A. McCauliff, *Burdens of Proof: Degrees of Belief, Quanta of Evidence, or Constitutional Guarantees?*, 35 VAND. L. REV. 1293, 1327–28 (1982) (summarizing a survey of federal judges).

[58] 565 U.S. 400, 430–31 (2012) (Alito, J., concurring).

[59] *See* McCauliff, *supra* note 57, at 1327; *see also* WAYNE R. LAFAVE ET AL., CRIMINAL PROCEDURE § 3.3(b) (5th ed. 2009).

[60] Jouvenal, *supra* note 54.

[61] Roviaro v. United States, 353 U.S. 53, 60–61 (1957). While *Roviaro* has been narrowly construed, it provides a possible basis, under the Compulsory Process Clause, for an argument that algorithms that determine a defendant's fate are discoverable. *See* Zathrina Zasell Gutierrez Perez, Note, *Piercing the Veil of Informant Confidentiality: The Role of In Camera Hearings in the* Roviaro *Determination*, 46 AM. CRIM. L. REV. 179, 202–13 (2009) (describing and critiquing federal circuit approaches to *Roviaro*).

who fit the profile are confronted or that people are targeted on a pre-specified basis, such as every fifth person (fairness). The precedent for this requirement comes from the Supreme Court's checkpoint jurisprudence, which requires that people stopped at roadblocks be selected on a neutral basis.[62] Unless this rule is followed, the biases that algorithms are meant to prevent will simply be reintroduced when police make the decision about whom to stop.

CONCLUSION

Further discussion of many of the ideas broached above can be found elsewhere.[63] The point to be emphasized here is that if risk assessment is a legitimate state exercise, it needs to be cabined by principles that demand that the methods used to implement it are legally germane, accurate, and fairly applied.

---

[62] *See, e.g.*, Delaware v. Prouse, 440 U.S. 648, 657 (1979) (distinguishing "[f]or Fourth Amendment purposes . . . between sporadic and random stops of individual vehicles making their way through city traffic and those stops occasioned by roadblocks where all vehicles are brought to a halt or to a near halt, and all are subjected to a show of the police power of the community."). *See generally* Christopher Slobogin, *Policing, Databases, and Surveillance: Five Regulatory Categories*, 12–13 (Nat'l Const. Ctr. White Paper Series, 2017), https://ssrn.com/abstract=2947948 [http://perma.cc/JMZ4-PVBR] (explaining the advantages of such an approach).

[63] *See* Slobogin, *supra* notes 21, 40, 56 & 62.