Washington University Law Review

Volume 96 | Issue 3

2018

Algorithmic Risk Assessments and the Double-Edged Sword of Youth

Megan T. Stevenson

Christopher Slobogin

ALGORITHMIC RISK ASSESSMENTS AND THE DOUBLE-EDGED SWORD OF YOUTH

MEGAN T. STEVENSON[*] & CHRISTOPHER SLOBOGIN[**]

INTRODUCTION

Risk assessment algorithms—statistical formulas that predict the likelihood a person will commit crime in the future—are used across the country to help make life-altering decisions in the criminal process, including setting bail, determining sentences, selecting probation conditions, and deciding parole.[1] Yet many of these instruments are "black-box" tools. The algorithms they use are secret, both to the sentencing authorities who rely on them and to the offender whose life is affected. The opaque nature of these tools raises numerous legal and ethical concerns. In this paper we argue that risk assessment algorithms obfuscate how certain factors, usually considered mitigating by sentencing authorities, can lead to higher risk scores and thus inappropriately inflate sentences. We illustrate this phenomenon through one of its most dramatic manifestations: The role of age in risk assessment algorithms.[2]

When considered as a factor at sentencing, youthfulness can be a double-edged sword—it can both enhance risk and diminish blameworthiness. If either risk or culpability is the sole issue at sentencing, this potential conflict is avoided. But when, as is often the case, both risk and culpability are considered relevant to the sentence, the aggravating effect of youth should presumably be offset or perhaps eliminated entirely by its mitigating impact. If judges and parole authorities are fully informed of the conflicting roles youth plays in a particular case, they can engage in this balancing act as appropriate. However, when information about risk comes from a black-box

* Assistant Professor of Law, George Mason University.

** Milton Underwood Professor of Law, Vanderbilt University. The authors would like to thank Brandon Garrett, Issa Kohler-Hausman, Sandra Mayson, David Robinson, Hannah Jane Sassaman, and Nicholas Scurich for their comments on earlier drafts of this article.

1. PAMELA M. CASEY ET AL., NAT'L CTR. FOR STATE COURTS, USING OFFENDER RISK AND NEEDS ASSESSMENT INFORMATION AT SENTENCING (2011) (describing how magistrates, correctional officials and judges use actuarial risk assessment instruments in deciding whether to grant bail or parole and in determining offenders' "criminogenic needs" when imposing sentence).

2. Other factors that might be considered both risk-aggravating and blame-mitigating include mental illness, substance abuse and lack of education. *See infra* text accompanying notes 95–97.

algorithm, they are unlikely to know the extent to which the risk evaluation is influenced by the defendant's youthfulness. In such cases, their decisions about pretrial detention, sentence, or release may unknowingly give youth too much weight as an aggravator.

Further, even if the black box is opened and the risk assessment algorithm is made publicly available, the risk score may not be conveyed in a fully transparent manner. For instance, while judges may be told that an offender's youth is a risk factor, the relative weight of age in the overall score may not be fully explained or understood at the time of decision-making.[3] Unless the judge makes specific inquiries, she will not be informed of the variables that contributed most heavily to a particular defendant's risk score.

This decisional blindness is especially pernicious in light of the impression created by the labels associated with these instruments—"high risk" or "high risk of violence." Such labels not only convey information about the potential for recidivism. They are also suggestive of bad character, or at least a history of bad decision-making. In other words, these labels convey condemnation. Such condemnation might be appropriate for an individual who has earned the "high-risk" classification by committing multiple violent or ruthless acts. But it is not warranted for an individual who has earned that label largely because of his or her youth.

To ensure sentencers take this double-edged sword problem into account, risk assessment algorithms should be transparent about the factors that most heavily influence the score. Only in that way can courts and legislators engage in an explicit discussion about whether, and to what extent, young age should be considered a mitigator or an aggravator in fashioning criminal punishment.

In Part I, we discuss the tensions youthfulness generates in the post-conviction setting by introducing the double-edge sword phenomenon and the jurisprudence that has developed around it. In Part II, we present empirical evidence that shows how influential age is in the widely-used COMPAS Violent Recidivism Risk Score (VRRS) and in other common risk assessment tools. Specifically, we conduct a partial decomposition of the VRRS to show that age alone can explain almost 60% of its variation, substantially more than the contributions of criminal history, gender or race. Similar patterns are documented in other common risk scores. In Part III, we discuss how obfuscation of age's impact on the risk score improperly undermines consideration of youthfulness as a mitigating factor. We also discuss how the points we make about the role of youth might apply to a

3. *See infra* text accompanying notes 84–86.

number of other factors that are often used in structured risk assessments, including mental illness, substance abuse, and socio-economic factors. While our discussion centers on sentencing, the main argument is generally relevant to a broad range of settings in which risk assessments influence criminal justice outcomes.

I. THE ROLE OF YOUTH IN SENTENCING

Reliance on youth at sentencing can raise at least three issues that resonate with constitutional prohibitions. The first is whether basing a sentence in whole or in part on youth raises an equal protection claim. Some have argued that age classifications, like those based on gender and race, should be subject to heightened scrutiny.[4] The second arises from the notion, recently solidified by the Supreme Court into constitutional doctrine where race is involved, that punishment should not be based on status.[5] Although both of these issues require careful thought, neither is addressed here.

The focus of this article will instead be on a third issue: how youth is used in conflicting ways at sentencing, and the reasons—arguably also of constitutional magnitude—that such use is questionable. This focus requires an examination of the "double-edged sword" dilemma, and the ways in which courts have dealt with it. While the issues discussed in this article relate to both the juvenile justice system and the adult criminal justice system, we focus our analysis exclusively on the adult system and adult risk assessment tools. Thus, when we refer to youths, we are referring only to those in their late teens and early twenties who are initially prosecuted in adult court, and to those under the age of eighteen who have been transferred to the adult system.

A. Youth as a Double-Edged Sword

In the American legal system, youth is often a strong basis for leniency in punishment. The most obvious evidence of this stance is the fact that every state has established a juvenile court system that diverts young offenders away from the harshness of adult criminal justice.[6] In adult

4. *Cf.* Nina A. Cohn, *Rethinking the Constitutionality of Age Discrimination: A Challenge to a Decades-Old Consensus*, 44 U.C. DAVIS L. REV. 213, 231 (2010) (focusing on protections for the elderly).

5. *See* Buck v. Davis, 137 S.Ct. 759, 778 (2017).

6. Barry C. Feld, *Juvenile Justice*, *in* 1 REFORMING CRIMINAL JUSTICE: INTRODUCTION AND CRIMINALIZATION 329, 330 (Erik Luna ed., 2017).

sentencing regimes as well, youth is often treated as a mitigator, which in some situations is even constitutionally required. In *Roper v. Simmons*,[7] the Supreme Court held that the imposition of the death penalty on juvenile offenders who have been transferred to adult court violates the Eighth Amendment. Seven years later, in *Miller v. Alabama*,[8] it concluded that the Eighth Amendment also bars mandatory life without parole sentences for juvenile offenders. In both decisions, the Court emphasized the inverse relationship between adolescence and culpability. As Justice Kagan put it in *Miller*, "[b]ecause juveniles have diminished culpability and greater prospects for reform, . . . 'they are less deserving of the most severe punishments.'"[9] Subsequent developments have made clear that the thousands of minors who are transferred out of the juvenile system and subject to sentences short of the death penalty and life without parole are also explicitly encompassed by the mitigation rationale developed in the Supreme Court's decisions.[10]

The Supreme Court's decisions draw a bright line at the age of eighteen. But the rationale of *Roper* and its progeny clearly does not evaporate at that age. Influenced by those cases, some jurisdictions have recently expanded juvenile court jurisdiction beyond eighteen.[11] More importantly, well before *Roper*, most capital sentencing statutes treated young (post-adolescent) adulthood as a mitigating factor when deciding whether the death penalty is appropriate.[12] A similar practice has long existed in non-capital sentencing practice. For instance, the Federal Sentencing Guidelines state that "age (including youth) may be relevant in determining whether a departure is warranted, if considerations based on age, individually or in combination

7. 543 U.S. 551 (2005).

8. 567 U.S. 460 (2012).

9. 567 U.S. at 471 (quoting *Graham v. Florida*, 560 U.S. 48, 68 (2010)).

10. *See* Feld, *supra* note 6, at 382 (stating that "states annually try upward of 200,000 chronological juveniles as adults" and noting that the Court's kids-are-different jurisprudence applies to such offenders, although it provides them only "limited relief").

11. Merril Sobie, *The State of American Juvenile Justice*, ABA CRIMINAL JUSTICE MAGAZINE, Spring 2018, at 26, 27.

12. *See, e.g.*, OHIO REV. CODE ANN. § 2929.04(B)(4) (LexisNexis 2018); UTAH CODE ANN. § 76-3-207(4)(e) (LexisNexis 2018); *see generally*, Jeffrey Kirchmeier, *A Tear in the Eye of the Law: Mitigating Factors and the Progression Toward a Disease Theory of Criminal Justice*, 83 OR. L. REV. 631, 673–75 n.226 (2004) (listing relevant statutes and caselaw). *Cf.* Johnson v. Texas, 509 U.S. 350, 367 (1993) ("There is no dispute that a defendant's youth is a relevant mitigating circumstance that must be within the effective reach of a capital sentencing jury if death sentence is to meet [constitutional requirements].").

with other offender characteristics, are present to an unusual degree and distinguish the case from the typical cases covered by the guidelines."[13] State regimes are often even more explicit about making post-adolescent youth a mitigating consideration at sentencing.[14]

At the same time, intuition suggests—and research indicates[15]—that youthfulness can also be an *aggravating* factor. The Supreme Court has often observed that juveniles are less "deterrable" than adults.[16] Although that language is meant to support the Court's conclusion that youth is a mitigating factor, it also recognizes that young people tend to be more impulsive, less risk averse, more easily influenced by peers, and less constrained by "stakes in life"—all of which tend to increase the likelihood that young people will engage in criminal activity.[17] If the goal is to prevent crime via incapacitation, one might argue that it is logical to incarcerate youths through the years of peak criminal activity.

As a legal matter, these facts give rise to the familiar "double-edged sword" problem.[18] Depending upon the purpose of punishment at issue, the same factor can be seen as either mitigating or aggravating. Because they diminish culpability and control, factors like youth and mental disability may well be mitigating from a retributive or deterrence perspective. From an incapacitative perspective, however, they may be aggravating factors, because they enhance risk.

Sentencing practices reflect this tension. In a meta-analysis of approximately 60 studies examining the effect of age on sentences, Jaejong Wu and Cassia Spohn found that 40% of the studies reported a positive relationship between these two variables (i.e., older offenders received longer sentences), 57% reported a negative relationship, and 3% reported

13. U.S. SENTENCING GUIDELINES MANUAL § 5H1.1 (U.S. SENTENCING COMM'N 2016).

14. For instance, a number of states have "youthful offender" provisions that call for more lenient treatment of offenders tried in adult court who are under a certain age (e.g., 21 or 25). *See, e.g.*, FLA. STAT. § 958.04; GA. CODE. ANN. § 42-7-2; N.J. STAT. ANN. § 30:4-148 (West 2018).

15. *See, e.g.*, John Monahan, Jennifer Skeem & Christopher Lowencamp, *Age and Risk Assessment, and Sanctioning: Overestimating the Old, Underestimating the Young* (2017), at http://www.ssrn.com/abstract=2973503 [https://perma.cc/FS24-7D94].

16. *See, e.g.*, Roper v. Simmons, 543 U.S. 551, 571 (2005) ("the same characteristics that render juveniles less culpable than adults suggest as well that juveniles will be less susceptible to deterrence."); Graham v. Florida, 560 U.S. 48, 72 (2010) (same); Miller v. Alabama, 567 U.S. 460, 471 (2012) (same).

17. *See generally* Christopher Slobogin & Mark R. Fondacaro, Juveniles at Risk: A Plea for Preventive Justice 19–28 (2011) (describing the research).

18. See, e.g., Rabindranath Ramana, Living and Dying with a Double-Edge Sword: Mental Health Evidence in the Tenth Circuit's Capital Cases, 88 DENV. U. L. REV. 339 (2011).

no relationship at all.[19] Wu and Spohn concluded that federal courts are more likely to treat youth as a mitigator than state courts, and that northern courts are more likely than southern states to treat youth as an aggravator.[20] The obvious legal question that arises from these observations is whether this differential use of youth at sentencing is permissible.

B. A Double-Edged Jurisprudence

At the most fundamental level, the role of youth in sentencing should be consistent with the relevant jurisdiction's sentencing policy. Using youth as an aggravator might be considered impermissible in a sentencing regime that is driven largely by retributive goals.[21] However, youth might permissibly be considered both an aggravator and a mitigator in sentencing regimes that are based on amalgams of retribution, deterrence, reformation, and individual prevention goals.[22]

In the latter regimes, judges who want to take seriously the sentencing impact of youthful characteristics will have to exercise a considerable degree of judgement. Take, for instance, a case involving a highly impulsive youth; as the Supreme Court's cases indicate, impulsivity can be both aggravating and mitigating. Faced with this situation, a judge might decide that such an offender is high risk, and also conclude that his ability to reason rationally is not so impaired as to require leniency, thus leading to a sentence enhancement. Or consider an offender whose youthful desire to please others both increases his risk level and decreases his culpability. Here a judge might conclude that the risk posed by the offender's vulnerability to peer pressure is relatively minimal and in any event treatable in the community, and that this fact, together with the offender's unformed character, permits a more lenient sentence.

Carrying out this nuanced type of analysis is difficult and subjective. Furthermore, it may be problematic on constitutional grounds. One of the reasons that people with intellectual disability are exempted from the death

19. Jaejong Wu & Cassia Spohn, Does an Offender's Age Have an Effect on Sentence Length?: A Meta-Analytic Review, 20 CRIM. JUST. POL'Y REV. 379 (2009).

20. However, Wu and Spohn also note that these disparities were less evident in those studies that did a better job controlling for other variables. *Id.* at 391.

21. *See generally* Richard Frase, *Punishment Purposes*, 58 STAN. L. REV. 67 (2005) (describing different sentencing systems, their rationales, and examples of each).

22. *Id.*

penalty is because the Supreme Court wanted to avoid the double-edged sword problem. In *Atkins v. Virginia*,[23] the Supreme Court expressed concern that intellectual disability "can be a two-edged sword that may enhance the likelihood that the aggravating factor of future dangerousness will be found by the jury."[24] In other words, the *Atkins* majority suggested that intellectual disability should be treated solely as a mitigator, and fashioned a holding that assured that result. The Court has voiced a similar sentiment in connection with mental illness. In *Zant v. Stephens* it favorably noted the fact that Georgia's capital sentencing statute had not "attached the 'aggravating' label to . . . conduct that actually should militate in favor of a lesser penalty, such as perhaps the defendant's mental illness."[25] In short, given the *Roper* line of cases establishing the mitigating relevance of youth, and the doubled-edge sword concerns expressed in *Atkins* and *Zant*, one could interpret current doctrine as suggesting that youth should *never* be treated as an aggravator.

There are arguments to the contrary, however, and the issue remains unresolved.[26] In the meantime, courts not only must grapple with the fact that youth can be both a risk-aggravator and a blame-mitigator, but also with an even more insidious double-edged sword problem. The problem is well-illustrated by *Huckaby v. Florida*,[27] which dealt with not with age, but with mental illness, a factor that, like age, can be relevant both to risk and to culpability. In *Huckaby*, a capital case, the trial court found that the defendant was suffering from a serious mental illness at the time of the offense, and thus had proven a statutory mitigator based on limited culpability.[28] However, the court still sentenced Huckaby to death, based in part on its conclusion that the murder had been committed in a heinous manner, a statutory aggravator under Florida law that posits that murders committed in a horrific manner are especially blameworthy.[29] On appeal, the Florida Supreme Court did not dispute the murder was carried out in a gruesome manner. Nonetheless, it reversed the sentence, finding that "[t]he heinous and atrocious manner in which this crime was perpetrated . . . [was] the direct consequence of [Huckaby's] mental illness, so far as the record reveals."[30] The higher court concluded, rightly in our view, that Huckaby's mental illness could not both mitigate blame and aggravate it.

23. 536 U.S. 304 (2002).
24. *Id.* at 321.
25. 462 U.S. 862, 885 (1983).
26. See CHRISTOPHER SLOBOGIN, MINDING JUSTICE: LAWS THAT DEPRIVE PEOPLE OF MENTAL DISABILITY OF LIFE AND LIBERTY 90–92 (2006).
27. 343 So. 2d 29 (Fla. 1977).
28. *Id.* at 34.
29. *Id.*
30. *Id.*

Similarly, even if youth can be both a risk-aggravator and a culpability-mitigator, presumably all agree that youth cannot be a culpability-*aggravator*. Youth is a life stage that everyone passes through and over which an individual has no control. There is nothing blameworthy about being young. Yet when judges are unaware of the impact that youthfulness plays in the risk score, there is a danger of engaging in the sort of illegitimate double-edged swordism demonstrated in *Huckaby*. That is because it is easy to associate a high risk label not just with dangerousness but also with bad character and condemnation. If that conflation occurs, youth can end up not only legitimately enhancing risk but also illegitimately enhancing *culpability*. We will return to this topic in Part III. For now, enough has been said to set the stage for a discussion of recent developments connected with risk assessment that have made the multiple potential roles of youth in sentencing an increasingly pressing issue.

II. YOUTH AS A RISK FACTOR: AN ANALYSIS OF THE COMPAS

In the past decade, a number of states have moved toward "evidence-based sentencing."[31] This type of sentencing instructs sentencing authorities to rely as much as possible on risk assessment and risk management instruments that structure the inquiry into an offender's dangerousness and treatability. Proponents of such tools argue that they can improve the ability to choose between incarceration, some intermediate community sanction, or complete release. Some of these instruments are strictly actuarial, meaning that they include only risk factors that are statistically-correlated with risk and produce numerical probability estimates of risk. Other instruments provide a more qualitative assessment, but still require a structured inquiry informed by the empirical literature on risk assessment and risk management.[32]

One of the more popular risk assessment instruments is the COMPAS (for Correctional Offender Management Profiling for Alternative Sanctions), developed by a company called Equivant (formerly Northpointe). The COMPAS algorithms were first developed in 1998 and

31. *See generally* Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 566–67 (2015).

32. For a detailed account of some of the more popular instruments, see RANDY K. OTTO & KEVIN S. DOUGLAS, HANDBOOK OF VIOLENCE RISK ASSESSMENT (2d ed. 2018).

have been revised several times since then.[33] Although originally designed to aid the departments of corrections in placing, managing, and treating offenders, it has since been used in other contexts, including pretrial decision-making, sentencing, and parole.[34]

The COMPAS has two primary risk models: General Recidivism Risk and Violent Recidivism Risk.[35] As the names suggest, the former model is a general predictor of future offending while the latter model predicts violent reoffending. While Equivant provides general information about the factors that are included, the exact algorithm is proprietary.

The COMPAS algorithm has stepped into the spotlight several times in recent years. It was integral to a 2016 case, *State v. Loomis*,[36] in which the petitioner argued that consideration of the COMPAS algorithm in determining his sentence violated his due process rights in three ways. First, the petitioner argued that because the algorithm is proprietary, defendants are unable to discern and challenge its scientific validity. Second, he contended that a sentence based on an algorithm derived from group data is not "individualized." And third, the petitioner argued that the COMPAS algorithm impermissibly takes gender into account.[37] Although the Wisconsin Supreme Court admonished lower courts to be aware of these concerns and avoid making the COMPAS score determinative,[38] it rejected all three challenges and affirmed Loomis's sentence of fifteen to twenty years, ten to fifteen years of which had been added by the trial court because of the risk he posed.[39]

Scholars and journalists have also paid special attention to the COMPAS algorithm. In 2017, the public advocacy organization ProPublica accused the COMPAS algorithm of being biased against black defendants.[40] This critique spawned an active discussion about the definitions of racial fairness in algorithms.[41] Similarly, a recent paper challenged the predictive

33. *Id.* at 2.
34. Erin Collins, *Punishing Risk*, GEO. L. J. (forthcoming May 2018) at 31.
35. Northpointe, *Practitioner's Guide to COMPAS*, August 17, 2012 at 1.
36. 881 N.W.2d 749 (Wisc. 2016).
37. *Id.* at 761–62 (upholding the use of the COMPAS tool in sentencing in Wisconsin)
38. *Id.* at 276.
39. *Id.* at 282.
40. *See* Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, at 1 (2017), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [https://perma.cc/REP7-T8G5] (describing the COMPAS as one of "two leading nationwide tools offered by commercial vendors").
41. *See e.g.* Jon Kleinberg et al., *Inherent Tradeoffs in the Fair Determination of Risk Scores*, PROCEEDINGS OF INNOVATIONS IN COMPUTER SCIENCE (November 17, 2016),

superiority of the COMPAS algorithm by claiming to show that a group of random online survey respondents given only 7 data points could predict recidivism as well as the COMPAS algorithm.[42]

Such events have generated considerable popular interest in the COMPAS algorithm. In this article, we open still another line of inquiry about the instrument, by examining the relative explanatory power of factors used in the COMPAS algorithm. While ProPublica's critique focused on the role of race in COMPAS, and the *Loomis* case focused on the COMPAS's use of gender, we examine the role that age plays in that instrument, and find that age contributes much more explanatory power than either of these factors. Indeed, we find that age contributes substantially more to the sentencing recommendations produced by the COMPAS Violent Recidivism Risk Score than any other variable. After explicating that conclusion, we look at several similar instruments and find that COMPAS is not unique in heavily weighting age; many other risk assessment algorithms also lean substantially on youthfulness when predicting recidivism.

A. Age in the COMPAS Violent Recidivism Risk Score

Equivant has publicly stated that the factors included in the COMPAS Violence Recidivism Risk Score (VRRS) consist of age at assessment, age at first adjudication, the History of Violence Scale, the History of Noncompliance Scale, and the Vocational Educational Scale.[43] The company says nothing about the extent to which each factor contributes to these scores, or how the various scales are constructed.

While the exact algorithm that constitutes the VRRS is secret, its construction can be at least partially deciphered through reverse engineering. As used here, reverse engineering refers to a process of evaluating the importance of a particular factor within a risk score by

https://arxiv.org/pdf/1609.05807.pdf [https://perma.cc/VET3-YEW7] (showing that, except in very specialized circumstances, achieving equal false positive rates for two groups with different base rates for criminal activity would require a risk tool to assign the same risk classification to different levels of actual risk across the two groups).

42. Julia Dressel & Hany Farid, *The Accuracy, Fairness and Limits of Predicting Recidivism*, SCIENCE ADVANCES 1, 1 (2018).

43. *See* William Dieterich et. al., COMPAS Risk Scales: Demonstrating Accuracy and Predictive Parity, Performance of the COMPAS Risk Scales in Broward County, Northpointe Inc., Research Department, Jul. 8, 2016, at 5–6.

determining how much of the variation in the score can be explained by that factor.[44] A complete reverse-engineering would allow explanation of 100% of the variation in VRRS: all of the factors that contribute to the score, as well as the weights on these factors, would be known. Here, given the inability to access all the relevant data on which the COMPAS relies, we are only able to partially reverse-engineer the algorithm. But, with that caveat, our model explains 72% of the variation in VRRS, and allows us to identify factors that contribute substantially to the overall score.

We conducted our analysis of the VRRS using a data set of risk scores acquired by ProPublica as part of its study on racial bias in risk algorithms in Florida.[45] Through a public records request, ProPublica received two years' worth (2013–2014) of COMPAS risk scores from the Sherriff's Office in Broward County.[46] Since COMPAS is predominantly used to determine pretrial custody in that jurisdiction, scores generated at other stages—such as parole and probation determinations—were dropped from the analysis.[47] Using name and date of birth, the risk scores were matched to public criminal records from the Broward County Clerk's Office Website, jail records from Broward County's Sherriff's Office, and public incarceration records from the Florida Department of Corrections website.[48] The data provided by ProPublica also contained information on each defendant's age (in years), race, gender, current charge, degree of current charge, number of prior arrests, and number of juvenile felony, misdemeanor, and "other" arrests. We supplemented these data with publicly available information on prior incarceration from the Florida Department of Corrections website.[49] Incarceration data was matched to the ProPublica data using first name, last name, and birthdate.[50] Thus the final data set tracks criminal history, demographics, and COMPAS risk scores for all defendants who received a COMPAS VRRS as part of pretrial processing between 2013 and 2014 in Broward County, Florida.

Table 1 describes the data. There are 4020 total cases, initiated on dates spanning the beginning of 2013 to the end of 2014. The average age of the

44. It is worth noting that this is a different empirical exercise than is often conducted with risk assessments; we are not trying determine how well the different factors predict recidivism, but rather how well they predict the COMPAS risk score.

45. Larson et al., *supra* note 40. The data can be downloaded from https://github.com/propublica/compas-analysis.

46. *Id*.

47. *Id*.

48. *Id*.

49. These data can be downloaded at http://www.dc.state.fl.us/pub/obis_request.html; the data used in this paper was downloaded in October 2015.

50. The incarceration data included both a dummy that is equal to one if the defendant has a prior incarceration in the state of Florida, as well as dummy variables for the specific security level of the facility.

defendants in the sample was 35; 36% of the defendants were white, 48% were black, and 9% were Hispanic; 79% of the defendants were male, and 40% faced felony charges.

Table 1. Description of the Data

Variable	Mean	SD	Min	Max	Variable	%	SD	Min	Max
Age	35.7	12.07	18	83	*White*	36	0.48	0	1
Juvenile felony arrests	0.04	0.43	0	20	*Black*	48	0.5	0	1
Juvenile misd. arrests	0.07	0.4	0	8	*Hispanic*	9	0.28	0	1
Juvenile other arrests	0.08	0.41	0	7	*Male*	79	0.41	0	1
Number of prior arrests	2.44	3.95	0	38	*Felony*	40	0.49	0	1
VRRS[51]	2.44	0.86	0.37	5.18	*Prior Incarceration*	7	0.25	0	1

Note: This table contains descriptive statistics for the 4,020 cases used in the analysis.

Using this data set, we examined the effect of seven different factors on COMPAS VRRS scores: age (our variable of interest),[52] current charges,

51. The raw risk score in the data is negatively signed. To avoid the confusion associated with a negatively signed risk score, we added five so that all scores are positive.

52. The age factor includes dummies for each age, rounded to the nearest year. The current charges factor includes dummies for the exact charge as well as dummies for the degree of the charge. The juvenile criminal history factor is as described in the next sentence in the text, and the race factor includes dummies for being black, white or Hispanic (mutually exclusive). The number of prior arrests factor includes dummies for the exact number of prior arrests. The prior incarceration factor includes a dummy

juvenile criminal history, race, number of prior arrests,[53] prior incarceration, and gender. Some of these factors consist of a number of variables. For instance, the juvenile criminal history factor contains all the variables that pertain to juvenile justice: the number of juvenile felony arrests, juvenile misdemeanor arrests, and "other" juvenile arrests (probably consisting of juvenile-specific offenses, such as curfew violation).[54] We then used several metrics to evaluate the extent to which each of these factors contribute to the risk score,[55] all relating to a statistic called the adjusted R^2.[56] Specifically, we looked at each factor's *individual* explanatory power (which does not take into account how the factors might interact),[57] *marginal* explanatory power (which does look at this interaction),[58] and *overall* explanatory power.[59]

Figure 1 shows the *individual* explanatory power of the seven factors we considered. As the figure indicates, age has substantially more individual explanatory power than any of the other factors. In fact, age alone explains 57% of the variation in VRRS.

---

for whether or not the defendant had a prior period of incarceration in Florida, as well as dummies for the security level of the custody (close, community, medium, minimum). The gender variable is binary.

53. The data provide information on the number of prior "counts" but do not specify what this word means. Here, we defer to Northpointe's response to ProPublica, in which they refer to these as prior arrests. Deiterich, *supra* note 43, at 6 (stating that ProPublica "did include age and number of prior arrests in the data they posted").

54. More specifically, the juvenile justice factor includes a fully saturated set of dummy variables for each number of juvenile felony, misdemeanor and "other" arrests.

55. We use the raw risk score, as provided in the file "compas-scores-raw.csv," which is matched to the other files using name, date of birth, and date that the risk score was completed. Following the ProPublica analysis, cases where the COMPAS score was taken more than 30 days from the date of arrest (723 in total) were dropped.

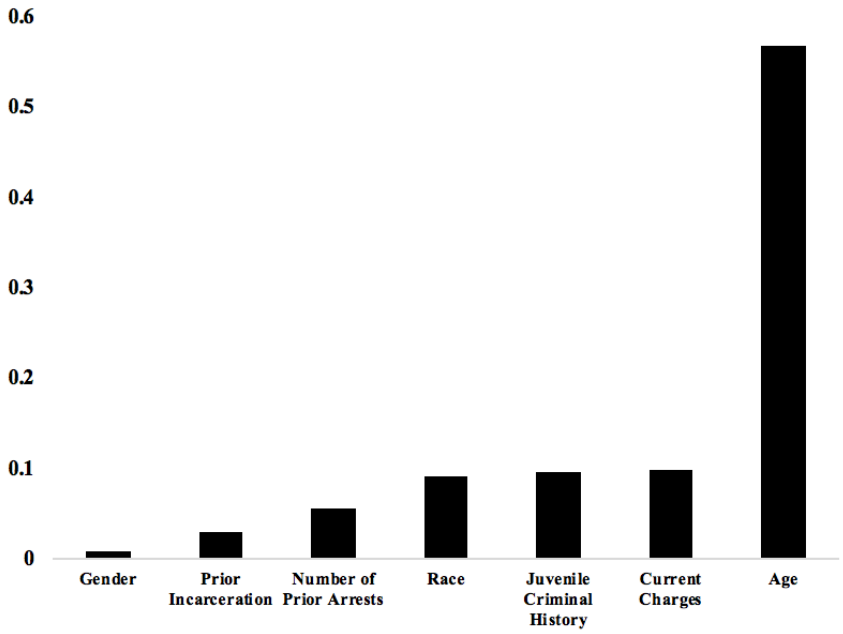
56. For a formal definition of adjusted R^2, see Jeffrey M. Wooldridge, INTRODUCTORY ECONOMETRICS A MODERN APPROACH 202 (5 ed. 2014). The R^2 is a statistical measure of how much of the total variation in the risk score is explained by that factor; the adjusted R^2 accounts for the number of variables in that factor. We focus on the R^2 instead of coefficient magnitudes for several reasons. First, coefficients can be large and statistically significant while still explaining only a relatively small portion of the variation. Second, some of the factors we consider consist of multiple variables; in fact, we chose to fully saturate the regressions, so as to avoid parametric assumptions about how, say, the number of prior arrests affects the score. A coefficient-based analysis would be hard to interpret. We could have considered the F statistic in a test of joint significance on multiple coefficients, but the R^2 is a more intuitive and easily explainable metric

57. The individual explanatory power of a factor refers to the adjusted R^2 in a linear regression of the VRRS on that factor individually. The individual explanatory power of the juvenile justice factor, for instance, includes not only the explanatory power of juvenile justice alone, but also some of the explanatory power of the factors with which it is correlated.

58. The marginal explanatory power differs from the individual explanatory power in that it represents the amount of explanatory power one factor adds after all the other factors have been accounted for. This second measure is defined as the difference between the adjusted R^2s in two linear regressions: a regression of the VRRS on all factors *including* the factor under consideration, and a regression of the VRRS on all factors *except* the factor under consideration.

59. The overall explanatory power is the adjusted R^2 from a regression of the VRRS on all factors or on various subsets of factors.

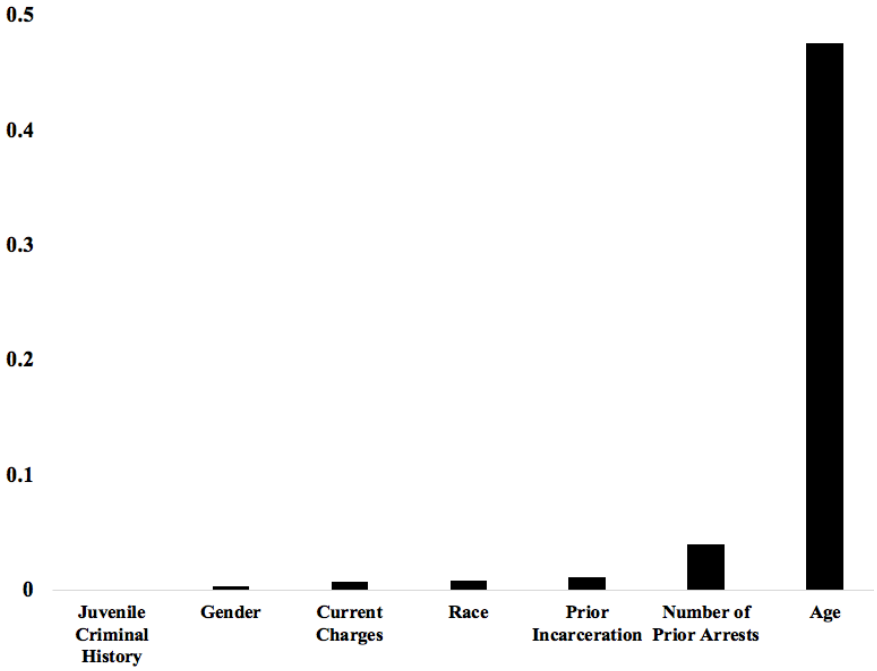
Figure 1: Individual Explanatory Powers of Various Factors on the COMPAS Violent Recidivism Score



Note: This figure shows the adjusted R^2 in a regression of the raw Violent Recidivism Risk Score on each of the factors listed above. Each regression includes only one factor at a time, although the factors may consist of multiple variables. For instance, the individual explanatory power of age is the adjusted R^2 in a regression of the VRRS on a set of 63 dummy variables for each year of age found in the data. The adjustment to the R^2 accounts for the number of regressors included in the model.

Figure 2 shows the *marginal* explanatory power of each factor on the VRRS. Once again, age has substantially more explanatory power than any other factor under consideration. Even after accounting for criminal history, current charge, age, gender, and race, age explains 48% of the total variation. All of the other factors have a marginal explanatory power of 5% or less.

Figure 2: Marginal Explanatory Powers of Various Factors on the COMPAS Violent Recidivism Score



Note: This figure shows the difference in the adjusted R^2 between two regressions: a regression of the raw VRRS on all of the factors and a regression of VRRS on all factors except the one listed in the leftmost column. Thus, the marginal explanatory power of a factor is how much explanatory power is contributed after all other factors have been accounted for. The adjustment to the R^2 accounts for the number of regressors included in the model.

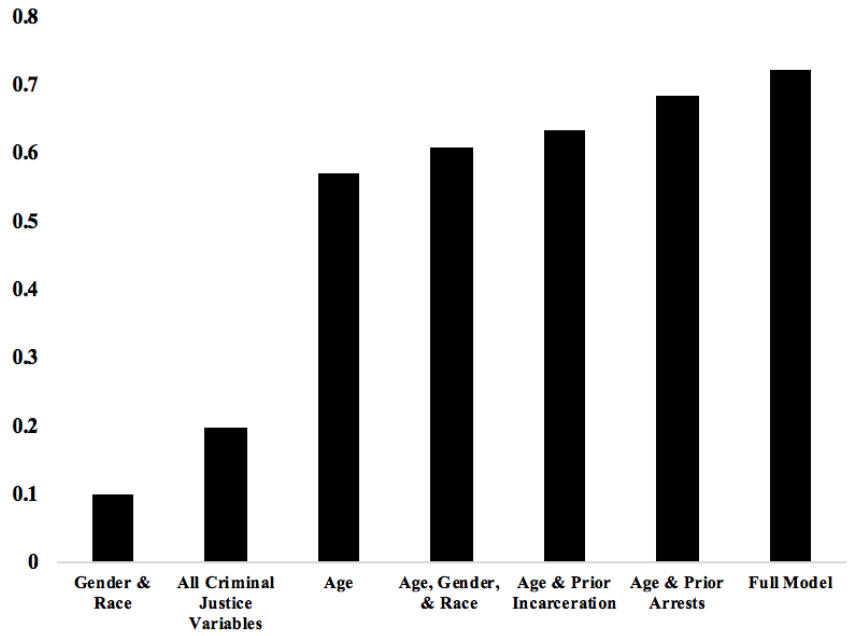
In considering these findings, it must be recognized that the high marginal explanatory power of age could be due to the fact that age is correlated with other inputs to the VRRS that are not available in our data. While, as explained more fully below, the seven factors we were able to investigate account for a substantial portion of the total variation in the risk score, they do not account for everything. For instance, if we had been able to control for employment status, which is likely correlated with age, the marginal explanatory power of age might have been less.

Does this diminish the argument that age is an important factor in the VRRS? Probably not. Regardless of whether age influences the risk score directly, through its inclusion in the algorithm, or indirectly, through its

correlation with other factors such as employment status, the fact remains that young people have much higher risk scores. Functionally, it makes little difference whether age affects the risk score by direct inclusion in the algorithm, or instead through age-proxies like lower employment.

Figure 3 shows the *overall* explanatory power of seven different models. The full model, which contains all seven factors, has the most explanatory power, explaining 72% of the variation in the VRRS. A very close second, however, is a model that contains only two factors: age and the number of prior arrests. This model can explain 68% of the total variation. Figure 3 also shows that age remains a potent factor even when combined with factors other than prior arrests. Finally, as we already indicated in Figure 1 (and display again here), even on its own age explains 57% of the variation, while the model that contains only criminal history variables (current charge, juvenile justice, prior arrests and prior incarceration) only explains about 20% of the variation, and the model containing gender and race explains only 10% of the variation.

Figure 3: The Overall Explanatory Power of Various Models

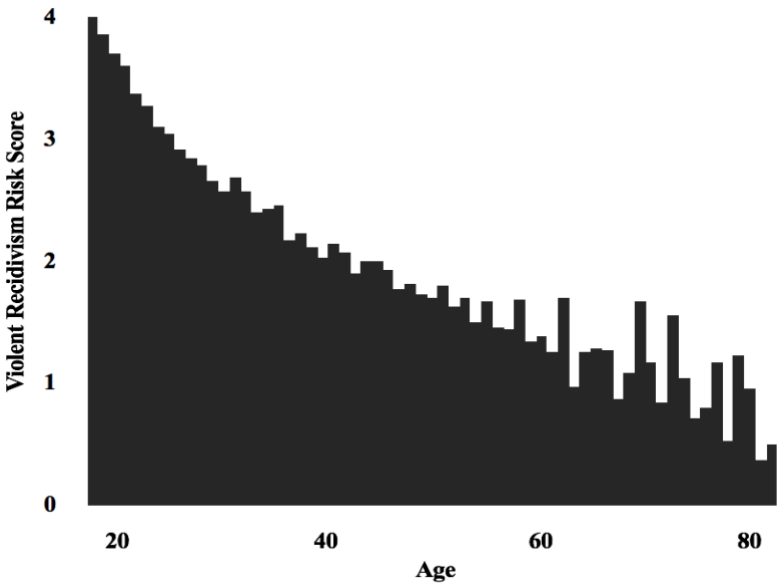


Note: This figure shows the adjusted R^2 of various models: one that includes all factors, and six that include only a subset of factors. The

adjustment to the R^2 accounts for the number of regressors included in the model.

As a way of bringing home the importance of age to risk assessment, Figure 4 shows the average risk score for defendants of each age.[60] There is a strong inverse relationship between age and the VRRS. The average risk score declines as age increases, and this decline is particularly steep between the ages of eighteen and thirty. Eighteen year old defendants have risk scores that are, on average, twice as high as forty-year-old defendants. We also conducted a partial reverse-engineer of the COMPAS General Recidivism Risk Score (GRRS). We found that age is still a predictor, although the magnitude of influence is less than in the Violent Recidivism Risk Score. On the GRRS, age has individual explanatory power of 27% and marginal explanatory power of 23%.

Figure 4: The Age and Risk Score Relationship



60. The raw risk score in the data is negatively signed, where zero is the highest risk and negative 5 is the lowest risk. To avoid the confusion associated with a negatively signed risk score, we added five points to each score so that all scores are positive.

In sum, our analysis of the COMPAS shows that age is one of the most important factors in the instrument. Ideally, one would not have to reverse engineer the COMPAS to arrive at this conclusion. But we had to do so given its developer's unwillingness to share its algorithm.[61]

B. Age in Other Risk Assessment Tools

The COMPAS algorithms are not the only risk assessment instruments in which age plays a large role. Among risk assessment tools that make their algorithm publicly available, many others place a heavy weight on age. Table 2 demonstrates this conclusion with respect to eight such instruments (some of which are more commonly used to aid pretrial detention decisions rather than sentencing determinations) by comparing the treatment of age with the treatment of criminal history.

The table clearly shows that, in each case, age is as heavily weighted or more heavily weighted than the comparative criminal history measure. Column 3 of the table shows the difference in scores between an eighteen-year-old and a fifty-year-old who are similar in all other respects. Column 4 shows the number of points that various criminal history measures add to the defendant's score. As a comparison of these two columns indicates, the influence of age is either identical to or greater than the influence of criminal history. For instance, in the PCRA, which is the risk assessment used at sentencing in the federal system, being eighteen adds two points to the risk score, the same number of points that are added for having three to six prior arrests. Among Virginia's three instruments, age adds between four and six more points than criminal history. In short, an assessment relying on these instruments is heavily dependent on age.[62]

---

61. Equivant does acknowledge that age "carries a lot of weight in the Violent Recidivism Risk Score calculation" and that "if you are young, unemployed, and have an early age-at-first arrest and a history of supervision failure, you could score a medium or high on the Violence Risk Scale even though you never had a violent offense arrest." *Practitioner's Guide to COMPAS*, *supra* note 35, at 25–26.

62. It should be noted, however, that age is not included in all risk assessment algorithms. The LSI-R does not list age among its inputs, although many of its inputs, such as employment status, are likely correlated with age. *See* WASHINGTON STATE INSTITUTE FOR PUBLIC POLICY, SEX OFFENDER SENTENCING IN WASHINGTON STATE: PREDICTING RECIDIVISM BASED ON THE LSI-R (2006) (a list of individual LSI-R items in the technical appendix does not mention age). Neither the Ohio nor the Indiana Risk Assessment Community Supervision tools (used in sentencing) include age as a direct input. *See* EDWARD LATESSA ET. AL., CREATION AND VALIDATION OF THE OHIO RISK ASSESSMENT SYSTEM, FINAL REPORT 51–53 (2009); University of Cincinnati, INDIANA RISK ASSESSMENT SYSTEM 2-4, 2-5, 2-6 (2010). The Ohio and Indiana Risk Assessment tools used to determine correctional facility

Table 2. Age in Various Risk Assessment Instruments

Risk Assessment	Type of Instrument	Score for Being 18	Score for Prior Criminal History	Prior Criminal History Specifics
PCRA[63]	Federal post-conviction	2	2	Three to six prior arrests[64]
Static-99[65]	Risk of sex offending, male	2	1	Any number of prior violent (but non-sexual) convictions[66]
PTRA[67]	Federal pretrial	2	2	Five or more prior felony convictions[68]
VRAI - Fraud, Larceny[69]	Virginia – Sentencing	21	15	Three or more prior adult felony convictions[70]
VRAI - Drug[71]	Virginia – Sentencing	8	5	Three prior adult felony convictions[72]
VRAI - Sex Offender[73]	Virginia – Sentencing	12	8	Two or more prior felony arrests and zero-three prior misdemeanor arrests[74]
PSA[75]	Arnold Foundation's Public Safety Assessment - pretrial	2	2	Three or more prior violent convictions[76]

placement and reentry do include age, but the weights associated with age are relatively low. LATESSA ET AL., *supra*, at 56, 60; UNIVERSITY OF CINCINNATI, *supra*, at 3-1, 4-1.

63. Jennifer L. Skeem & Christopher T. Lowenkamp, *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 689 (2016).

64. *Id*.

65. Amy Phenix et al., Static-99R Coding Rules 46 (rev. 2016).

66. *Id.* at 58.

67. OFFICE OF PROBATION AND PRETRIAL SERV., FEDERAL PRETRIAL RISK ASSESSMENT (PTRA) USER'S MANUAL AND SCORING GUIDE 1 (2010), https://www.pretrial.org/download/risk-assessment/Federal%20Pretrial%20Risk%20Assessment%20Instrument%20(2010).pdf [https://perma.cc/J58K-P663].

68. *Id*.

69. VA. CRIMINAL SENTENCING COMM'N, FRAUD WORKSHEET (2017), www.vcsc.virginia.gov/worksheets_2017/fraud.pdf [https://perma.cc/55BA-E4ZF]; VA. CRIMINAL SENTENCING COMM'N, LARCENY WORKSHEET (2017), https://www.vcsc.virginia.gov/worksheets_2017/Larceny.pdf [https://perma.cc/U97M-6JMQ].

70. *See* sources cited *supra* note 69.

71. VA. CRIMINAL SENTENCING COMM'N, DRUG SCHEDULE I/II WORKSHEET (2017), http://www.vcsc.virginia.gov/worksheets_2017/SchI_II.pdf [https://perma.cc/U6RC-7WRP].

72. *Id*.

73. VA. CRIMINAL SENTENCING COMM'N, RAPE WORKSHEET (2017), http://www.vcsc.virginia.gov/worksheets_2017/Rape_.pdf [https://perma.cc/98UM-7BTQ].

74. *Id*.

75. Arnold Found., Public Safety Assessment, Risk Factors and Formula 3 (2016).

76. *Id*.

The information depicted in Table 2 is publicly available for those who are willing to invest the time into searching it out. But, importantly, it is generally not made salient to the judge at the time of use. Unless the judge is numerically skilled and has memorized the weights on different factors, we expect that there is a less-than-perfect understanding of the impact different factors like age have on the risk score. That lack of transparency is a significant concern, especially at sentencing, where relative youth is generally supposed to be a mitigating factor.

III. THE HIDDEN DANGERS OF THE DOUBLE-EDGED SWORD CONUNDRUM

Young age is a highly influential factor in modern risk assessment instruments. But that influence is not immediately apparent from the risk scores that legal decision-makers use to make pretrial detention, sentencing, and release determinations. This lack of transparency means that decision-makers may not realize the extent to which a high risk score is based on a factor that they may consider mitigating.

The problem created by the opacity of these instruments goes much deeper than that, however. As Part I demonstrated, while a particular factor, such as youth or mental illness, might logically be considered both a culpability-mitigator and a risk-aggravator, it should never function as both a culpability-mitigator and culpability-*aggravator*. Yet, given the ambiguous nature of a high risk finding, that illegitimate conflation is precisely what might happen if the role of youth in risk assessment is not made obvious. After explaining the problem further, we explore ways of dealing with it.

A. Risk Assessment as Character Judgement

Consider two individuals. James is an eighteen-year-old male facing a marijuana possession charge. He has a pending charge (also for marijuana possession) and has failed to appear in court on one occasion. Carl is forty-years-old, is facing a charge for aggravated assault, has two prior convictions (one for armed robbery and one for selling cocaine), and has spent two years in prison. As we have shown, it is very possible that James and Carl will receive the same score on the COMPAS VRRS assessment instrument. But while James's high risk evaluation would largely be due to his age, Carl's high risk evaluation would instead largely be due to his prior violent convictions. As implied by Table 4, violent crime rates drop considerably with age, so eighteen-year-olds with no history of violence

may actually have the same statistical risk of violence as forty-year-olds with multiple prior violent convictions.

Is it important that the judge be aware of which factors contribute to high-risk labels such as these? In particular, is it important that the judge know that James has received the "high risk of violent recidivism" label primarily because of his age? We think so, in part because of the way risk assessments are often perceived and used by judges and other decision-makers.

Risk assessments convey more than information about statistical risk; whether intended or not, they are also often interpreted as statements about character. For instance, one routinely finds linkage of the defendant's dangerousness with an assessment of his or her character in judicial decisions.[77] Indeed, the Supreme Court itself has made the connection, when it stated in *Deck v. Missouri* that evidence of "danger to the community . . . almost inevitably affects adversely the jury's perception of the character of the defendant."[78] This type of pronouncement reflects the intuition that a statement about someone's propensity for committing a violent offense can easily be interpreted as a statement about that person's intrinsic worth.

In *Deck*, the Supreme Court went on to say that "character and propensities of the defendant are part of a 'unique, individualized judgment regarding the punishment that a particular person deserves.'"[79] Many commentators have likewise observed that character is closely related to blameworthiness and desert. For instance, Professor Peter Arenella has argued that there is no means of judging persons charged with crime except through assessing their character.[80] Professor James Whitman has contended that consideration of character "makes it possible to consider the full spectrum of information about individual blameworthiness, including

77. Sherron v. State, 2017 WL 6521705 *2 (Ind. App. 2017) (upholding a sentence because of a finding of the offender's "character as being 'predatory, disturbing, dangerous'"); State v. Bell, 33 A.3d 167, 181 (Conn. 2011) (speaking of the "public's interest in protecting itself from dangerous criminals and in imposing a fair sentence on the basis of the defendant's history and character"); Casillas v. State, 941 N.E.2d 572 *3 (Ind. App. 2011) (unpublished table decision) (upholding sentence because of the offender's "violent and recklessly dangerous character"); State v. Day, 551 N.W.2d 871 *1 (Wisc. 1996) (unpublished table decision) (describing the offender's sentence as based on the offender's "crime, character, and dangerousness").

78. 544 U.S. 622, 633 (2005) ("[t]he appearance of the offender during the penalty phase in shackles . . . almost inevitably implies to a jury, as a matter of common sense, that court authorities consider the offender a danger to the community—often a statutory aggravator and nearly always a relevant factor in jury decisionmaking . . . [and] almost inevitably affects adversely the jury's perception of the character of the defendant.").

79. *Id.* (quoting Zant v. Stephens, 462 U.S. 862, 900 (1983).

80. Peter Arenella, *Character, Choice and Moral Agency: The Relevance of Character to Our Moral Culpability Judgments*, 7 SOC. PHIL. & POL'Y 59 (1990).

both dangerousness and deservingness."[81] And Professor Kyron Huigens has developed an aretaic theory of punishment that views the criminal justice system primarily as a means of judging and improving character.[82]

If a risk assessment is interpreted as an assessment of character, and if judges and other sentencing authorities believe that character is closely related to blameworthiness, then inclusion of youth in a risk assessment tool becomes particularly problematic. Return to the examples of Carl, the forty-year-old assault offender with a history of armed robbery, and James, the young marijuana user. Carl's choice to commit serious crimes in the past might reasonably be seen as an expression of bad character that can be used in aggravation whether the focus of sentencing is risk *or* blameworthiness. In contrast, it obviously is not correct to say that James's young age alone is indicative of bad character. Nonetheless, that is precisely what sentencing authorities signify when they allow their sentencing decisions to be determined by algorithms that place so much weight on youthfulness.

Furthermore, if the sentencing authority does in fact enhance James's sentence based on character inferences drawn from the "high risk of violent recidivism" label, it is engaging in the worse sort of double-edged swordism. As we noted above, as a logical matter youth may be both a mitigator and an aggravator if its use in aggravation is focused *solely* on future behavior. But if youth is instead used as an indicator of bad character and not just as an indicator of high risk, it has been associated with moral condemnation. This is, of course, in direct contradiction to the traditional position—reinforced by the Supreme Court's kids-are-different jurisprudence—that youth diminishes culpability.

This double-edged sword problem is exacerbated when the risk assessment is based on an instrument like the COMPAS, because of that instrument's lack of transparency. In using such a tool, judges might *unknowingly and unintentionally* use youth as a blame-aggravator. A judge who is aware of the defendant's youth may consider it partially excusing, due to the reduced culpability of youth. But if that same offender is denominated "high-risk" and the judge, unaware that this label is heavily influenced by the defendant's youthful age, interprets it as a statement of bad character, then youthfulness unwittingly contributes simultaneously to moral condemnation. That result is both illogical and unacceptable.

81. James Q. Whitman, *The Case for Penal Modernism: Beyond Utility and Desert*, 1 CRITICAL ANALYSIS L. 143, 178 (2014).

82. Kyron Huigens, *The Dead End of Deterrence, and Beyond*, 41 WM. & MARY L. REV. 943, 1022–34 (2000)

B. The Need for Transparency

This is not the first paper to call for greater transparency in risk assessment use. Other authors who have written about algorithmic decision-making have argued that biases in data collection and analysis cannot be exposed unless courts or some other supervisory authority has access to the underlying code.[83] To these contentions, we add a new argument for transparency in the criminal justice setting, one based on the potential for unexamined risk assessments to produce results that are inconsistent with the avowed purposes of criminal punishment.

To understand the need for transparency, consider the risk assessment process in Virginia, one of the leading states in using risk assessment instruments at sentencing. As Table 2's analysis indicates, youth plays a significant role in the assessment tools Virginia judges use. Yet these judges may not realize that fact; indeed, a recent survey of Virginia judges indicated that only 29% reported being "very familiar" with their risk assessment tool, and 22% were either "unfamiliar" or only "slightly familiar" with it.[84] Given these statistics, it is unlikely that most Virginia judges fully comprehend the role that youthfulness plays in their risk instrument. The matter is exacerbated by the fact that risk scores in Virginia are paired with explicit sentencing-directives; for example, the state's sentencing guidelines recommend that judges divert non-violent offenders who are rated "low risk."[85] While judges retain discretion as to whether to follow such directives, they are not apprised of how the risk scores are produced unless they specifically ask for such information.[86]

Furthermore, to avoid the problems we have identified, simply making public the factors included in the risk assessment will not be enough. That limited type of transparency will make it too easy for judges to conclude

83. *See generally*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 4 (2015); Lilian Edwards & Michael Veale, *Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?* 16 IEEE Security & Privacy 46 (2018); Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018); Elizabeth E. Joh, *The Undue Influence of Surveillance Technology Companies on Policing*, 92 N.Y.U. L. REV. ONLINE 101 (2017).

84. JOHN MONAHAN ET. AL., NONVIOLENT RISK ASSESSMENT IN VIRGINIA SENTENCING, REPORT 2: A SURVEY OF CIRCUIT COURT JUDGES 7 (2018) https://content.law.virginia.edu/system/files/news/spr18/Judges%20sentencing%20survey%20March%201.pdf [https://perma.cc/7PSN-5V8R].

85. Kevin R. Reitz, *"Risk Discretion" at Sentencing*, 30 FED. SENT'G REP. 68, 70 (2017).

86. Further, their decisions are non-appealable. Luttrell v. Commonwealth, 592 S.E.2d 752, 755 (Va. App. 2004) ("[A] trial judge's failure to correctly apply the sentencing guidelines 'shall not be reviewable on appeal or the basis of any other post-conviction relief'.").

that, since age is one among many risk factors, its inclusion in the instrument raises no important issues.[87] In contrast, if judges are made fully aware of how influential age is in the risk score—how, for instance, it accounts for almost 60% of variation in some risk assessment algorithms—their reaction is likely to be very different. With this additional information, judges will be in a better position to balance the mitigating and the aggravating aspects of youth.

Fortunately, this situation can be remedied in large part by providing legal actors with relevant and easily interpretable information about how specific risk factors are weighted.[88] This type of transparency may be possible even for instruments that are proprietary, if the relevant company is willing to surrender limited control over its code. For instance, the risk score could be conveyed along with information about only the most important factors. The judge would be informed "The defendant's risk score is A, and the three most influential contributing factors are X, Y, and Z."[89]

To return to James and Carl, the judge might be informed that the most important factors in the risk score for James are age, the pending charge, and the prior failure to appear. For Carl, the judge would be told that the paramount factors are the current violent charge, the prior violent convictions, and the prior incarceration. Such information would likely go a long way towards alleviating the potential conflation of blame and risk that we have identified.

One could, of course, ask for more. For instance, courts could be told the precise number of points associated with the most important risk factors. For algorithms that have a relatively small number of inputs—a dozen or less, say—it would be tractable to inform the judge of how much *every* characteristic adds to the risk score for a particular defendant. At the same time, for ease of use the factors that are most heavily weighted could be highlighted.

87. An example of this phenomenon is found in State v. Loomis, 881 N.W.2d 749, 767 (Wisc. 2016) (upholding a sentence based in part on a risk assessment that included gender because the trial court considered "multiple factors").

88. The tools we discuss here are not the product of artificial intelligence or machine, which pose greater obstacles to interpretability. *See generally*, Michael L. Rich, *Machine Learning, Automated Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871, 883–86 (2016). But even here some level of transparency is possible. *See generally* Joshua A. Kroll, et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017).

89. There are multiple ways of defining which factors are "most influential" in determining a risk score. The most effective method likely depends on the specifics of the tool and the context.

Private companies might not be willing to surrender their entire code in this way, however.[90] And some may be unwilling even to reveal the most important risk factors at issue. Arguments can and have been made—based on the right of confrontation or due process—that courts should be able to force them to do so.[91] We will not canvass those arguments here; suffice it to say that we believe they are convincing. If such arguments fail, states that wish to continue using risk assessments could and should develop their own algorithms.[92]

The type of transparency that would ameliorate the concerns addressed in this paper requires neither a change in risk instrument used, nor a radical departure in practice. As a technical matter, the necessary procedure is relatively easy to implement. Legal barriers may be more substantial, but should not stand in the way if accurate and fair sentencing is the goal.

C. Other Risk Factors

Some risk factors are likely to be aggravating from either a backward-looking retributive perspective or a forward-looking risk perspective. The most obvious example is criminal history. Repeated criminal conduct suggests both that a person is more culpable and higher risk.[93] Other, more contentious examples in this vein might include membership in a gang and failure to complete a previously-imposed rehabilitation program.[94] While transparency is probably always preferred, it is not as important with respect to these factors because the high-risk label and its evocation of bad character will not improperly hide or misuse a blame-mitigator.

A different response is necessary, however, with risk factors that, like youth, are best described as mitigators when viewed from a retributive perspective. If the courts have explicitly made such a determination, as they seem to have done with mental illness, then the double-edged sword problem we have described in connection with youth arises once again.

---

90. For instance, Equivant, the creator of COMPAS, has refused to do so.

91. *See* sources cited *supra* note 83.

92. Pennsylvania, for instance, has done so. *See* Rhys Hester, *The Pennsylvania Experience with Risk Assessment Sentencing*, *in* RISK AND RETRIBUTION: THE ETHICS AND CONSEQUENCES OF PREDICTIVE SENTENCING (Jan W. de Keijsser, Julian V. Roberts & Jesper Ryberg, eds., forthcoming).

93. Julian Roberts & Richard Frase, *Predictive Sentencing: The Problematic Role of Prior Record Enhancements*, in RISK AND RETRIBUTION, *supra* note 92.

94. *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) (describing the HCR-20 risk assessment instrument, which includes treatment failures as a risk factor).

Thus, when such factors are treated as risk-enhancing in a risk assessment instrument, that fact must be made known to the sentencing judge so that its contribution to the risk appraisal can be balanced by its mitigating impact, and not obscured by a generic risk score.

Other risk factors are more difficult to characterize. Consider substance abuse. While substance abuse is usually considered a significant risk factor,[95] it could also easily be viewed as a blame-mitigator.[96] A number of other potential risk factors—including a history of unemployment, a lack of education, and one's residence—could be seen as blame-mitigators as well,[97] depending on the sentencing authority's view of how much control individuals have over such circumstances and how much they contributed to the offender's crime. This is a large issue which we will not tackle here.[98] But if these types of factors are included in a risk assessment tool, the difficulty of categorizing them argues for the same type of transparency that is clearly required when youth is the risk factor at issue.

CONCLUSION

The use of risk assessment tools in criminal justice is expanding rapidly. In this article we do not take a position on whether that is a good or bad development. Rather, assuming that risk assessment will be a significant feature in many sentencing regimes, we have argued that factors that are meant to mitigate blame—clearly youth, likely mental illness, and possibly many more—can only be treated that way if sentencing judges are given full information about the extent to which risk assessments instruments rely on them.

95. See *id*.; see also Seena Fazel et al., Prediction of Violent Reoffending on Release from Prison: Derivation and External Validation of a Scalable Tool, 3 LANCET PSYCHIATRY 535, 537 (2016), both of which designate substance abuse as a risk factor.

96. For instance, although the number has dwindled in recent years, several states recognize a voluntary intoxication defense. *See* WAYNE R. LAFAVE, CRIMINAL Law 498–501 (5th ed. 2010).

97. Several authors have argued, for instance, for a "rotten social background" defense. *See, e.g.*, Richard Delgado, *The Wretched of the Earth*, 2 ALA. C.R. & C.L. L. REV. 2, 22 (2011) ("[U]ntil we loosen the bonds that inhibit upward mobility, we have no business punishing the wretched of the earth who find themselves trapped in the bottom layers of society and, predictably, grow up without many controls or options."); Andrew Taslitz, *The Rule of Criminal Law: Why Courts and Legislatures Ignore Richard Delgado's Rotten Social Background*, 2 ALA. C.R. & C.L. L. REV 80, 129 (2011) ("The rotten social background defense calls us to a more inclusive, realistic, compassionate, and equal form of moral and legal rule.").

98. One of us has argued that unless these types of factors contribute significantly to the risk score, they should not be included in risk assessment instruments. Christopher Slobogin, *Principles of Risk Assessment: Sentencing and Policing*, 15 OHIO ST. J. CRIM. L. 583, 592–93 (2018).