6 Ethical Implications of Neurobiologically Informed Risk Assessment for Criminal Justice Decisions: A Case for Pragmatism

Eyal Aharoni, Sara Abdulla, Corey H. Allen, and Thomas Nadelhoffer

6.1 Introduction

The criminal justice system has a problem: it is tasked with protecting society from dangerous offenders, but it cannot reliably predict who will and who will not reoffend. Given that its predictions are imperfect, efforts to deter or incapacitate dangerous offenders will sometimes encroach on the rights of non-dangerous individuals, and conversely, efforts to protect the non-dangerous will sometimes harbor some "bad apples." To the extent that predictive errors can be reduced, our society could become safer and more just.

One potential way to reduce predictive error is to develop better risk assessment tools. In the criminal justice context, risk assessment refers to any technique that generates probabilistic predictions about the ostensible offender's behavior—such as the probability of reoffending, relapsing, or responding to treatment—by querying information about their attributes or circumstances. Risk assessment is employed throughout the criminal justice pipeline, from intake and sentencing to release, mainly to inform decisions about supervision (such as security level assignments or eligibility for parole) and treatment. It is also used in civil commitment settings such as involuntary hospitalization for individuals found to be dangerous but not guilty by reason of insanity. Traditionally, offender risk estimates were determined using unstructured clinical judgment by trained forensic psychologists—a subjective technique that has since been shown to perform little better than chance (Monahan & Silver, 2003). Recently, structured actuarial risk assessment techniques have improved the accuracy of offender placement decisions by quantifying the offender's degree of fit with a known validation sample. As a result, more than twenty states now

require their courts to use these statistical tools in service of offender sentencing and placement decisions (Starr, 2014).

Despite the increasing popularity of actuarial risk assessment, its use in legal settings has been the subject of much scholarly debate. When is it justified to use statistical models based on group data to determine legal outcomes for individual offenders? Risk assessment research suggests that a variety of biological indicators of risk—such as genes and brain function—may carry some degree of statistical weight in predictions of risk. Are some statistical indicators of risk more ethically problematic than others? Given the rapid advances in risk assessment research, it is imperative to anticipate and clarify when, if ever, statistical indicators should inform justice-related risk assessment.

Many ethicists suggest that all actuarial risk assessment is too problematic for use in justice settings, cautioning about violations of beneficence (e.g., unjustified harm to offenders), justice (e.g., unfair distribution of sanctions), and respect for persons (e.g., unjustified restrictions of the offender's freedom or exposure of his private mental life), among other problems. In this analysis, we examine some of the main ethical concerns, with an eye toward so-called neuroprediction technologies—that is, the use of neurobiological markers in offender risk assessment. We attempt to glean insight into some of the normative intuitions surrounding such technologies by evaluating their strengths and weaknesses relative to other alternative strategies. In this way, our analysis is fundamentally contrastive.

We conclude that while some uses of actuarial risk assessment might potentially violate individual rights to beneficence, justice, and respect for persons, these problems arise not just for evidence-based tools but for any decision procedure that society adopts to protect the public safety and civil rights of its members by trying to identify potentially dangerous individuals. We therefore attempt to shift the debate from *whether* actuarial risk assessment is justified to *when* (see also Specker et al., 2018). We argue that appeals to individual rights alone are not sufficient to distinguish between the ethically appropriate and inappropriate applications of actuarial risk assessment. Any principled attempt to evaluate the appropriateness of risk tools must, *for each application*, evaluate its unique costs *relative* to its benefits and *relative* to traditional clinical approaches (i.e., the status quo). When applied to various uses by the law, we find that actuarial risk assessment often fares better on key ethical criteria than traditional clinical methods.

Broadly, we find that appreciation of these relational contexts in which risk assessments are solicited carries the potential to clarify the underlying ethical concerns about risk assessment.

6.2 Legal Applications of Risk Assessment

Before we discuss the internal characteristics of common risk assessment techniques, it is important to understand the variety of ways in which these techniques can be, and often are, employed in legal settings.

- After a person is arrested, risk information might be useful to prosecutors making charging decisions. Prosecutors have wide discretion in such decisions. So, if a defendant is perceived to be particularly dangerous, a prosecutor might seek to file more charges or more severe charges.
- Bail decisions may also be informed by risk information. A defendant assessed to be a low flight risk may be offered bail or probation instead of jail time (the default). Likewise, bail fines can be increased or decreased depending on whether a defendant has a high or low risk score.
- Risk assessments can also inform whether the individual is placed on probation or instead sentenced to prison. Risk scores are not typically used to influence the *length* of a prison sentence, but may be used in pre-sentencing reports and could be considered relevant to actual sentencing decisions—for example, in Texas, capital sentencing hinges in part on a determination of future dangerousness (Tex. Code Crim. Proc., Art. 37.071).
- Within a facility, risk information can determine security level assignments: Authorities may assign low, medium, or high supervision based on the offender's recent behavior in custody, not as punishment but as a protective measure.
- A prison inmate could be released to parole prior to his sentence end date on grounds of a low risk assessment score. Similar uses are invoked when deciding to incarcerate a defendant via criminal versus civil commitment.
- Treatment interventions can be based on risk assessment. In some states, such as Florida, judges can mandate treatments such as anti-androgen therapy (i.e., "chemical castration") to certain sex offenders in addition to their original sentence based on their risk of reoffending (Fla. § 794.0235). Risk assessment can also inform decisions to administer treatment (typically noninvasive treatment) as a sentencing diversion.

In all of these ways, risk assessment is integral to many types of justice decisions either in service of or in opposition to the offender's private interests (see also Baum & Savulescu, 2013; Coppola, 2018; Tonry, 2014).

6.3 Risk Assessment Methods

6.3.1 Traditional Risk Assessment

Two broad types of risk assessment are employed within the legal system. Traditional risk assessment techniques rely heavily on unstructured clinical judgment to assess risk of antisocial behavior. Here, the driving factors in a clinician's determination of risk are the clinicians' (1) specific training, (2) experience with similar cases in the past, and (3) subjective assessment of the offender based on the case file, the crime, and (sometimes) an interview with the offender. The poor predictive accuracy of traditional clinical techniques has been demonstrated repeatedly. As Monahan (1981/1995) pointed out more than thirty years ago:

> It would be fair to conclude that the "best" clinical research currently in existence indicates that *psychiatrists and psychologists are accurate in no more than one out of three predictions of violent behavior over a several-year period among institutionalized populations that had both committed violence in the past (and thus had high base rates for it) and who were diagnosed as mentally ill*. (pp. 48–49; emphasis in original)

In an effort to explain why clinical risk assessment is so inaccurate and unreliable, Monahan (1981/1995, p. 32) identified what he takes to be "the four most common blind spots" of the clinical method: (1) a lack of specificity in defining the criteria being used, (2) a failure to take statistical base rates into consideration, (3) a reliance on bogus correlations, and (4) a failure to account for situational and environmental factors.

Given that clinical assessment is inherently subjective, it is unsurprising that the resulting predictions of future dangerousness are so unreliable (Lidz, Mulvey, & Gardner, 1993; Monahan, Brodsky, & Shan, 1981). In light of these types of problems, some commentators have even gone so far as to suggest that relying on clinical risk assessment for the purposes of the law is tantamount to "flipping coins in the courtroom" (Ennis & Litwack, 1974). Worse still, the unguided and intuitive nature of the process also makes it possible for the biases and prejudices of clinicians to influence their assessments. As Monahan (1981/1995) points out, "It is important to distinguish between the factors clinicians believe they are using—correctly

or incorrectly—to predict violent behavior and the factors that actually appear to influence their decisions" (p. 31). In summarizing the primary weaknesses of clinical risk assessment, Krauss and Sales (2001) make the following observation:

> In addition to relying on cognitive biases and heuristics that affect the judgments of ordinary people under conditions of uncertainty . . . mental health practitioners have been found to poorly combine information, use irrelevant information, and inappropriately vary the information they use in formulating predictions for an individual. Worse, their propensity for gathering excessive and irrelevant information also likely leads mental health practitioners to have greater confidence in their conclusions than is warranted. (p. 279; references omitted)

As a result, clinical risk assessments, perhaps unsurprisingly, tend not to be consistent from one mental health professional to the next.

Indeed, the American Psychiatric Association (APA) filed an amicus brief in the landmark case *Barefoot v. Estelle* (463 U.S. 883 [1983]). On their view, "the large body of research in this area indicates that, even under the best of conditions, psychiatric predictions of long-term future dangerousness are wrong in at least two out of every three cases." Because the ability of psychiatrists to reliably predict future dangerousness was "unproved," the APA claimed that "psychiatric testimony on future dangerousness impermissibly distorts the fact-finding process in capital cases." The court nevertheless dismissed the concerns of the APA, with the following reasoning:

> Neither petitioner nor the Association suggests that psychiatrists are always wrong with respect to future dangerousness, only most of the time. Yet the submission is that this category of testimony should be excised entirely from all trials. We are unconvinced, however, at least as of now, that the adversary process cannot be trusted to sort out the reliable from the unreliable evidence and opinion about future dangerousness, particularly when the convicted felon has the opportunity to present his own side of the case.[1]

In short, the court concluded that the adversarial nature of our legal system was sufficient to address the concerns raised by the petitioner with respect to the general unreliability of predictions of future dangerousness. Thus, the decision in *Barefoot* established that violence risk assessments are admissible in capital sentencing. The court has subsequently permitted these assessments in lower stake settings as well. The current state of the law is such that clinical risk assessments, despite their serious problems, are admissible and widely used in the criminal law. Given that risk assessment is here to

stay for the foreseeable future, the key question is: Can we improve upon the kinds of unstructured assessments that have traditionally been used?

6.3.2 Actuarial Risk Assessment

Whereas clinical risk assessment employs intuitive and subjective methods, actuarial risk assessment employs explicit, statistical, algorithmic methods. These methods work by matching features of the individual with an existing data set with known risk levels. These features can include clinical assessment of major mental illness, antisocial behavior, social support, and substance abuse history, to name a few. However, developers of actuarial risk assessment go beyond the clinical assessments by quantifying and validating exactly which types of data are to be considered in the prediction (Monahan, 1981/1995, p. 64). These quantitative approaches have demonstrated marked improvement in predictive accuracy (Dawes, Faust, & Meehl, 1989; Steadman et al., 2000; Banks et al., 2004), and meta-analyses have confirmed the superiority of evidence-based methods to clinical ones (Grove & Meehl, 1996).

Consider, for instance, the Violence Risk Appraisal Guide (VRAG), a twelve-item actuarial tool for predicting future dangerousness (Quinsey et al., 1998). VRAG uses the following static and historical predictor variables that have each been shown to confer risk for violence:

1. whether the offender lived with both biological parents as a child
2. whether the offender exhibited maladjustment in elementary school
3. whether the offender has history of alcohol problems
4. the offender's marital status
5. the offender's nonviolent criminal history prior to offense
6. whether the offender has failed on prior conditional release
7. the offender's age at the time of offense
8. the nature of the victim's injuries (e.g., death, hospitalization, treated and released, or slight)
9. whether the victim was a female
10. whether the offender meets the DSM criteria for a personality disorder
11. whether the offender meets the DSM criteria for schizophrenia
12. the offender's score on the Psychopathy Checklist Revised (PCL-R).

Not all predictor variables carry the same weight, as some variables (e.g., PCL-R score) are more predictive than others (e.g., alcohol problems). Offenders are categorized into the following risk levels: low, medium, and high risk. Each level consists of three more fine-grained "bins." Quinsey and colleagues (1998) found that "VRAG score was positively related to the probability of at least one violent reoffense, to the severity of the reoffenses that occurred, and to the speed with which violent reoffenses occurred" (p. 163). Since its development, VRAG has been tested within diverse populations and across a wide variety of settings.

One common way of testing the validity of a predictive model uses a receiver operating characteristic (ROC) analysis. The main reason for implementing a ROC analysis is that this method of representing predictive accuracy "is independent of the base rate of violence in the study sample" (Monahan et al., 2001, p. 95). The statistic used to summarize a ROC analysis is called the area under the curve (AUC), that is, "the probability that a randomly selected violent patient will have been assessed . . . as being in a higher risk group than a randomly selected nonviolent patient" (Monahan et al., 2001, p. 95). When one performs a ROC analysis, the AUC varies from 0.5—that is, accuracy is no better than chance—to 1.00—that is, perfect accuracy. Across multiple studies, Quinsey and colleagues (1998) found that the AUC for VRAG ranges from 0.73 to 0.77.[2] While this isn't ideal, it outperforms the clinical models we discussed earlier by a wide margin.

Given the increased predictive accuracy of models such as VRAG, some jurisdictions have begun to utilize actuarial risk assessment in decisions regarding criminal sentencing, supervised release, and the treatment and support programs required during supervision. Assessing individual offenders' risk levels, matching them to risk-appropriate programs, and targeting their unique criminal risk needs increases success, reduces antisocial behavior, and enhances public safety (Andrews, 2006; Aos, Miller, & Drake, 2006; MacKenzie, 2006; Taxman, 2002). Indeed, the traditional "one-size-fits-all" group treatment programs that ignore individual risk level can have damaging effects. One traditional group treatment program, for instance, reduced recidivism for high-risk offenders by more than 25 percent but led to an increase in reincarceration of low-risk offenders (Latessa, Lovins, & Smith, 2010). Because of the behavioral benefits of risk-appropriate treatment, some states have even begun to require the use of actuarial risk assessment

in legal decisions (e.g., Arkansas SB750; Public Safety Improvement Act 2011, Pub. L. No. SB750).

Though actuarial predictions represent an improvement over clinical judgment, they are by no means perfect. However, every additional percentage increase in risk assessment accuracy could reduce victimizations by targeting high-risk offenders while at the same time reducing our heavy reliance on mass incarceration by diverting low-risk offenders to treatment programs outside of prisons. For these reasons, scientists are beginning to examine whether the inclusion of neurobiological and/or genetic markers can improve predictive models based solely on non-biological evidence.

6.3.3 Neurobiologically Informed Risk Assessment

Neuroprediction refers to the use of neuroscientific measures to characterize biological markers of human behavior that increase the ability to classify particular behavioral outcomes accurately (e.g., Aharoni, Vincent, et al., 2013; Berns & Moore, 2012; Camchong, Stenger, & Fein, 2012; Delfin et al., 2019; Janes et al., 2010; Just et al., 2017; Pardini et al., 2014; Paulus, Tapert, & Schuckit, 2005; Sinha & Li, 2007; Steele, Fink, et al., 2014; Steele, Claus, et al., 2015; Steele, Maurer, et al., 2017). Currently, their use has been largely limited to research settings rather than legal settings.

One reason why courts have not used neuroprediction might be that neuroprediction and other biopredictive techniques have an unwholesome history. In the nineteenth century, it was believed that the morphology of a person's skull revealed information about their personality and behavior—a now-obsolete practice known as phrenology. Phrenology also inspired theories about criminal behavior, including one theory by Italian physician Cesare Lombroso that criminal offenders can be identified from a unique physiological signature, and that their criminal dispositions are biologically inherited, inferior, and unchangeable (Ellwood, 1912; Lombroso, 2006; Verplaetse, 2009). On this essentialist and deterministic view, many offenders are "born criminal"—a view that had a problematic relationship with eugenics and scientific racism. In the wake of Nazism, the explanatory pendulum swung toward situational models for human behavior that didn't appeal to biomarkers. These theories assumed that people are born into this world as blank slates and that any differences between them are the result of environmental and social factors such as poor parenting or poverty (e.g., Heider, 1958; for a critical analysis, see Pinker, 2016).

It is now widely held that human behavior is the result of a complex interplay of biological, psychological, and social factors aptly named the biopsychosocial model (Raine, 2013). For example, early lead exposure has been shown to increase the risk of later violent and aggressive behavior by impairing the development of brain areas such as the anterior cingulate cortex (ACC; Cecil et al., 2008; Liu et al., 2011). Several biomarkers have been developed to estimate the presence of lead poisoning (Sanders et al., 2009). However, while brain dysfunction caused by lead exposure represents a biomarker, it is also shaped by environmental factors as well, since people who live in homes or neighborhoods that have high levels of lead tend to be more socioeconomically disadvantaged. So, lead exposure represents an insult at two levels: the biological and the social. As such, it helps highlight the interdependency between the environment and the brain. The notion that environmental factors influence biological states that in turn affect behavior should not be surprising, given that all stimulus-based learning and behavior is mediated by the brain, the development of which is driven in large part by the coding of our genes along with nutrition and sometimes injury and disease.[3] One litmus test of the biopsychosocial model is whether explanations of human behavior demonstrate greater predictive accuracy when the relevant biological, psychological, and social factors are all included in the model.

The hypothesis that predictions of antisocial behavior can be improved by including targeted brain metrics in risk models formed the basis of a series of peer-reviewed studies that together have lent support to this hypothesis. For example, in a sample of ninety-six adult offenders who engaged in an impulse control task while undergoing functional magnetic resonance imaging, brain activity within the ACC prospectively predicted being rearrested later (Aharoni, Vincent, et al., 2013). Offenders with relatively low ACC activity had roughly double the odds of getting rearrested for a violent or nonviolent crime as those with high activity, controlling for other known risk factors. Using advanced statistical techniques such as AUC, calibration, and discrimination, these neuropredictive models have demonstrated relatively high accuracy (e.g., the probability that a true positive was correctly classified exceeded 75 percent; Aharoni, Mallett, et al., 2014; for a review, see Poldrack et al., 2018). The findings comport with the ACC's known role in error processing and impulse control (e.g., Kerns et al., 2004).

In a test of the convergent validity of these methods, researchers have also examined the predictive utility of ACC-based models using machine learning techniques that attempt to classify predictive functional patterns from event-related potentials (ERPs) measured in members of the original sample of Aharoni and colleagues (Steele, Claus, et al., 2015). The best performing model included a brain component known as the P300, which was capable of predicting being rearrested with comparably high accuracy. Other studies have reported similar predictive effects using different models of brain function (Delfin et al., 2019; Kiehl et al., 2018; Pardini et al., 2014). Together, these studies provide preliminary evidence that brain factors such as ACC function may serve as candidate biomarker for antisocial behavior.

Finally, several studies of substance-dependent forensic populations have highlighted the potential value of neuroprediction for purposes of assessing the probability of drug relapse and treatment completion, which are themselves known predictors of the risk of reoffending (Camchong et al., 2012; Janes et al., 2010; Paulus et al., 2005; Sinha & Li, 2007; Steele, Fink, et al., 2014). For example, ERP components and functional connectivity associated with the ability to monitor response errors have separately predicted treatment completion with accuracy exceeding 80 percent (Steele, Fink, et al., 2014; Steele, Maurer, et al., 2017).

Neuroprediction is still very much a nascent research field. Much remains to be understood about how our biological makeup shapes our behavior. The observed predictive effects are not and will never be perfect. But as the science advances, models that ignore the influence of biology on behavior will inevitably be forced to compete with those that acknowledge such influence. So, it is critical to understand better the impact that biology has on antisocial behavior and to think more carefully about how such behavior should be managed by society.

6.4 Evaluating Common Objections to Actuarial Risk Assessment

Many of the concerns surrounding the usage of neuropredictive technologies apply more generally to all actuarial risk assessment methods. So, rather than focus more narrowly on neuroprediction, we will focus primarily on actuarial prediction (except when there are salient differences between the two). Given that actuarial risk assessment of all kinds has been the target of persistent objections, we think our approach is appropriate. In responding

to critics of actuarial prediction, we identify four common types of concerns: statistical concerns and concerns about potential violations of norms of beneficence, justice, and respect for persons.

6.4.1 Concerns about Statistics

Critics have decried the low predictive accuracy among actuarial risk assessment tools (e.g., Starr, 2014). But this claim is misleading, as meta-analyses have demonstrated that the predictive value of many actuarial risk assessment instruments is superior to the traditional methods that would otherwise be used (e.g., Grove & Meehl, 1996). So, while we should continue to strive to increase the power of actuarial risk assessment, it already represents a significant step forward, even if it falls short of some critics' idealized standards. But from a legal perspective, its objective accuracy is beside the point. The Supreme Court has already made it clear that risk assessment is legally relevant. Thus, the only remaining legal questions are which risk assessment techniques we should incorporate and for which purposes.

A related and perhaps more compelling criticism is that the predictive accuracy of actuarial risk assessment holds only when making predictions about groups rather than individuals. This is referred to as the group-to-individual (G2i) inference problem. Indeed, scientists and practitioners must exercise great caution when considering whether and when to render diagnostic judgments based on group data (see Campbell & Eastman, 2014; Faigman, Monahan, & Slobogin, 2014; Heilbrun, Douglas, & Yasuhara, 2009). Predictable differences between groups do not imply that predictions about individuals within those groups will be similarly accurate (Buckholtz & Meyer-Lindenberg, 2012).

However, there are confusions about this objection that warrant clarification. A strong form of this objection states that, statistically, risk assessments may not be applied to individuals in principle because they are derived from group data. This stance cannot be true. By definition, the classification accuracy statistics employed by at least some of these instruments specifically describe an *individual's* degree of fit with a group, given certain shared attributes (Dawes et al., 1989). Thus, at least in principle, it is statistically justifiable to use these actuarial risk assessment techniques to make predictions about individuals.

Moreover, even if it weren't justifiable, this should be at least as problematic for clinical judgment. Like actuarial risk assessment, clinical judgment

is an inductive process that relies on past observations. Whether explicitly or implicitly, clinicians reference attributes from past cases—either textbooks or personal experience—to inform their predictions. The assumption is that this offender will behave as similar offenders have behaved in the past. So, if the use of group-based inference renders actuarial risk assessment problematic on statistical grounds, then *all* risk assessment would thereby be problematized.

The weak form of the G2i objection states that statistical risk assessments are not yet accurate enough to be practicably applied to individual offenders, who might vary widely from the group average. However, this claim demands greater precision. In the medical field, the accuracy of some of the most common screening tools is often marginal. For example, the ten-year cumulative probability of a false-positive mammogram result (a diagnosis that cancer is present when, factually, it is not) from breast cancer screening may exceed 60 percent (Hubbard et al., 2011). Yet, it is part of the standard of care for individual patients. For a doctor screening for cancer, it makes sense to tolerate a lot of false-positives in exchange for a high true-positive rate because the purpose of the screening phase is usually not to deliver a major medical intervention but rather to gather more information. Here, the intended application matters. Likewise, in criminal law, it is important to scale the accuracy of the predictive tool, regardless of type, to the stakes of the decision. But given that risk assessment is already pervasive in the law, the question isn't whether risk assessment is accurate enough, but rather how to improve it and when to use it.

Given that risk assessment, despite its limitations, will continue to play a role in the law, it seems prudent to advocate for the most reliable and accurate methods available. Indeed, scholars have made a powerful case for why statistical tools that generate G2i inferences, despite their caveats, can be responsibly leveraged for risk assessment (Bedard, 2017; Faigman et al., 2014; Heilbrun et al., 2009). Therefore, when authors object to the use of actuarial risk assessment on grounds of problematic G2i inference, a charitable interpretation is that there is nothing inherently problematic about G2i inferences in principle, but its present-day accuracy levels fail to meet *normative* standards required by the particular applications of our justice system. In other words, the concern about G2i is not so much a scientific objection as it is a moral or legal one (Janus & Prentkey, 2003). These moral and legal objections are the subject of the next sections.

6.4.2 Concerns about Beneficence

One moral concern about prosecutorial uses of actuarial risk assessment is that classification errors could result in the gratuitous infliction of harm. For example, false-positive errors could support a longer sentence or revocation of privileges from a non-dangerous offender, and false-negative errors could result in an early release of a dangerous offender, placing other community members at risk of victimization. To avoid these errors, it may be tempting to invoke a simple rule to minimize unnecessary harm.

However, removing risk assessment does not reduce harm. Judges are obligated to make sanctioning decisions with consideration for public safety regardless of whether an actuarial risk assessment is employed. The only other alternative (clinical judgment) still yields classification errors that are likely to be substantially larger without actuarial risk assessment than with it (Monahan & Silver, 2003). So, revocation of actuarial risk assessment on grounds of its harmfulness perversely increases harm by increasing the number of low-risk offenders who will be needlessly sanctioned and also the number of community members who will become victims of a truly dangerous person who was misclassified.

Another concern is that actuarial risk assessment facilitates the erroneous view that offenders are incorrigible, depriving them of appropriate service opportunities and possibly triggering a self-fulfilling prophecy (Specker et al., 2018). Perceived and anticipated stigma has been shown to predict poorer social adjustment following offender release, and thus is an important factor when considering potential harms inherent to actuarial risk assessment (Moore, Stuewig, & Tangney, 2016). However, this is a problem for all forms of risk assessment, not just actuarial risk assessment. Even with traditional clinical assessment, offenders are labeled as high risk, low risk, and so on, which can be stigmatizing, depending on the context. In this respect, actuarial methods are once again superior to clinical methods, given that the former methods permit the definition of both dynamic (i.e., changeable) risk and protective factors, such as the presence of positive or negative peer influences, which help service providers to identify appropriate interventions based in part on established evidence of treatment responsivity. Because these factors can be formally built into the actuarial models, this is likely to result in less, not more, stigmatization compared to traditional tools.

6.4.3 Concerns about Justice

Perhaps the most common normative concern about offender risk assessment is that it violates principles of distributive justice or fairness. In this view, it is unfair to judge an individual based on group tendencies because classification errors are not randomly distributed across individuals. If classification errors are non-random, this could facilitate arbitrarily worse treatment for certain, often underprivileged, groups. Risk assessment using biomarkers may also medicalize criminality in a way that stigmatizes already marginalized groups (Berryessa, 2017; Coppola, 2018; Hall, Carter, & Yücel, 2014; Jurjako, Malatesti, & Brazil, 2018). In this view, we ought to judge people not as mere statistics but for who they are as individuals. But structured risk assessment tools, it is argued, erroneously target individuals based on group membership rather than their true dangerousness, thereby codifying and reinforcing our inherent social biases (e.g., Silver & Miller, 2002, p. 152; Starr, 2014, p. 807).

The argument that people have the right to be treated as individuals as opposed to members of a (usually underprivileged) group often invokes the Fourteenth Amendment's Equal Protection Clause, which mandates that the government cannot discriminate on the basis of differences that are not relevant to a legitimate governmental interest (Starr, 2014, p. 827). Here, it is suggested that by assuming that the individual is a member of a high-risk group, actuarial risk assessment deprives that individual of the presumption that their actions should be judged independently of the actions of other members of the groups to which they belong. They are said to be punished for what other members of their group do.

However, for this argument to be coherent, it cannot assume that individuals ought to be judged completely free of *any* group membership. That would imply complete subjectification of the law: an extreme and unfeasible requirement to apply different laws to each individual in accordance with some ostensible model of his or her "true" self. A more plausible interpretation of the fairness argument, then, is that actuarial risk assessment erroneously assumes that the individual is a member of the unfavorable higher-risk group instead of the more legally privileged lower-risk group.

Importantly, whether this group classification is actually erroneous depends on the accuracy of the prediction. If this prediction turns out to be correct, and the individual is actually of high risk, the individual forfeits the law's more favorable presumption (that they belong to the low-risk group).

However, even if the high-risk prediction turns out to be wrong, this does not mean the individual is entitled to be judged independently of all group membership. Rather, it means that he should be judged as a member of the more privileged lower-risk group. For this reason, we can't reject actuarial risk assessment on grounds that it treats individuals as members of a group because the only other coherent alternative assumes these individuals are just members of another group (see also Janus & Prentkey, 2003). The Supreme Court seemed to recognize this point when it argued that equal protection must "coexist" with the fact that most legislation engages in classification in a way that disadvantages some social groups (Hamilton, 2015).

The law has been fairly explicit about which group factors are appropriate targets of risk assessment and which are not. Under the Equal Protection Clause, the Supreme Court has delineated three levels of scrutiny of evidence: strict scrutiny, intermediate scrutiny, and the rational basis test—the lowest level of scrutiny. In risk assessment, race typically falls under strict scrutiny, gender under intermediate scrutiny, and age under the rational basis test (Monahan, 2014).[4] So, the use of race as a factor requires a much stronger justification than the use of age, for example, to be considered for admissibility. Why is race held to a higher standard? One reason might be that race has been the basis of ongoing discrimination in the U.S. In particular, African Americans are disproportionately represented in the criminal justice system due in part to arbitrary and unfair policing and adjudication practices, and so corrective measures are needed to offset this error. Another reason for holding racial status to a higher standard might be that race is not independently diagnostic of offending (e.g., although U.S. prison populations are disproportionately African American, most African Americans are not criminals). A third reason might be that unlike dynamic risk factors such as drug use, race is static: people cannot change it at will. However, all of these reasons also arguably apply to some lesser-scrutinized factors such as age. So, they do not easily explain the difference in standards.

Perhaps a better explanation for why race evidence is held to a higher standard than age is that whereas there are good theoretical reasons to expect that age (as a measure of maturation) might play a direct causal role in some criminal behavior, no respectable theory suggests that race itself causes crime. So, one key function of these legal standards could be to filter out evidence that is less causally relevant because the susceptibility of such factors to error and bias will tend to outweigh their probative value.

Conversely, the law's lower standard for age suggests at least a tacit recognition that while a given risk factor does not need to be independently diagnostic of recidivism to be relevant, it should at least bear a direct correspondence to a known causal contributor to recidivism (e.g., as age bears to maturation).

A related reason that age merits a lower evidentiary standard might be that it cannot easily be used as a marker for other persecuted groups. For example, knowing someone's age is not helpful in predicting their race. Conversely, risk factors such as race, zip code, and socioeconomic status could more easily be used to target already-persecuted groups disproportionately. The complication is that some of these factors—such as socioeconomic status—may also contain causally relevant, independently predictive information, while others—such as race—do not. In such cases, completely ignoring the role of such factors in actuarial models would only serve to obscure any biases in classification. A better solution would be to control for problematic factors (such as race) statistically so that any variance attributed to those factors can be subtracted from otherwise useful factors (such as socioeconomic status). To avoid misuse of surrogate factors, it is useful to code for those factors explicitly, not to pretend they don't exist (Aharoni, Anderson, et al., 2019).

Some scholars have suggested, on grounds of discrimination, that we shouldn't support any selection rule whose outcomes might disfavor certain social groups (e.g., Starr, 2014; Tonry, 2014). This argument seems to confuse at least two different definitions of discrimination: unfair *outcomes* and unfair *processes*. By analogy, consider the familiar social rule: "For sports team photos, taller people have to stand in the back row." For some sports, such a rule might incidentally disfavor a certain race (e.g., if African American team members happen to be taller than others on average), but whether this rule would be *discriminatory* in the legal sense is questionable because it is blind to race from a selection perspective (i.e., it's not optimized to remove opportunities for African Americans as a group). The racial disadvantage in this case is only a *corollary* of an otherwise useful selection factor, not as a direct *causally* relevant selection factor itself. This is important because causal selection factors will tend to have lower covariation with other extraneous factors than more distant corollaries do.

Arguably, discarding the height rule would introduce even greater discrimination against shorter people, including many with physical disabilities, whose faces could be even more obscured from the camera's view than

the taller people were. In this mundane analogy, we do not mean to minimize the gravity of the issue of racial bias in the justice system. We use this analogy only to obviate the point that inequalities in *outcome* are often inescapable, but they do not necessarily indicate a need for a system overhaul. Discarding all rules that result in unequal outcomes among vulnerable groups would mean eliminating all risk assessment and either releasing all dangerous offenders into the community or locking up all non-dangerous ones. The fallout from either of these scenarios would be no consolation for those aiming to achieve equal outcomes. A humbler goal, and one that the Equal Protection Clause seems to support, is equality in legal process (Hamilton, 2015; Weiss, 2017).

What about risk factors that are more clearly biological, such as measures of brain activity? In his examination of the status of biological risk factors for violence risk assessment within the context of the Equal Protection Clause, Monahan (2014) concludes that as long as the risk factors in question are not being used as a surrogate for a more strictly regulated factor, biological risk factors would likely be admissible under a lower level of scrutiny. We agree with this point, provided that the biological evidence is demonstrated to be causally relevant. The broader lesson here is that while caution is warranted in deciding which risk factors merit inclusion in risk assessment instruments, these are tractable obstacles whose solutions can advance the utility of actuarial risk assessment beyond that of traditional clinical approaches.

The question at hand is not whether actuarial risk assessment is imperfect. It undoubtedly is. The question is whether the use of particular actuarial risk factors is any more problematic than the other alternatives (see also Dawes et al., 1989; Douglas et al., 2017; Nadelhoffer et al., 2012). Admittedly, the problem becomes more difficult when relatively accurate selection criteria are not available. So, under these conditions, it might indeed be justified to forgo the broader mission in order to protect minority groups. Our point is not to advocate all uses of actuarial risk assessment uncritically but instead to suggest that their justifiability must be judged relative to the ways in which they are to be used and what the other alternative options are.

On the broader point that actuarial risk assessment could codify society's social biases (Starr, 2014), we agree that these instruments might do so. It might ensure that some amount of bias is present in the predictions. Surely, for these tools to be effective and ethical, developers should always aim to minimize bias. But codification of bias is not the biggest problem faced by

these tools. By design, actuarial risk assessment attempts to codify risk factors. This feature makes the process transparent and enables us to monitor and evaluate procedural mistakes and ultimately to minimize bias. For this reason, codification of bias can be understood more as a strength than as a flaw because it enables the assumptions of the predictive models to be explicit and subject to evaluation and revision.

That is not to say that all actuarial models will invariably codify bias. Whether and the degree to which a model is biased will depend on the predictor variables that it uses. For instance, while researchers could likely develop more accurate models if they included race as a predictor variable for violence—not because different races are naturally more or less violent but precisely because races are treated differently by the legal system—they virtually never do. The main reason is that if researchers used race as a predictor variable, then their actuarial models wouldn't be admissible for legal purposes. Do these models nevertheless use variables that are proxies for race, for example zip codes or socioeconomic status? Some do, but most of the main models do not (or at least not obviously so). Are some of the twelve items used by VRAG proxies for race? Here, again, it's possible, but it's not obvious either way. Yet, because VRAG uses specific variables transparently, we can look at these variables and determine whether some may be serving as surrogates for race. If so, we can remove these variables and see whether it affects the overall predictive validity of the model. If removing the item impacts accuracy, then we can and should have an open discussion concerning whether the trade-off is worth it. Depending on the legal context—for example, the death penalty versus bail—we might come down on one side of the issue or the other.

This level of transparency is not found in traditional clinical methods. As cognitive psychologists have long warned, when human beings make decisions in the absence of constraints, we overly rely on anecdotal information and discount true probability information (e.g., Kogut & Ritov, 2005). Importantly, we may still consult probability information, we just do it unconsciously, intuitively, and unchecked (Ryazanov et al., 2018), using our own flawed, selective memories. When it comes to unstructured decision making, we are *folk* statisticians (de Hevia & Spelke, 2009; Kahneman & Tversky, 1972; Kunda & Nisbett, 1986; Tajfel & Wilkes, 1963; Tversky & Kahneman, 1973). In light of this well-studied fact, actuarial risk assessment is no more problematic—indeed, it is less problematic—than

traditional risk assessment. If we dispense with structured risk information in justice decisions, assessment bias does not go away, it just exerts its nasty effects behind a curtain of opacity.

6.4.4 Concerns about Respect for Persons

Regardless of whether risk assessment makes the offender the target of harm or discrimination, one might object to its use on grounds that it violates a basic obligation of respect for persons, including a requirement to protect the offender's autonomy—his natural freedom or right to make his own choices—and the privacy of his mental life. After being convicted, there are many ways in which the state already restricts the offender's autonomy and privacy, but at least prior to conviction and sentencing, decisions about how he lives his life should be up to him. Undergoing a neurological assessment of that person's dangerousness could violate his autonomy and privacy and therefore should be subject to his consent, just as it would be with any other citizen. Many offenders, of course, would not provide such consent, and even if they did, their consent may not be valid because the conditions under which it is solicited could be highly coercive. Bioethicist Paul Wolpe (2009) articulated this concern when concluding that "[t]he skull should be designated as a domain of absolute privacy," even in times of national security, on the grounds that bureaucratic intrusion into people's brains is to take away their "final freedom."

This concern about respect for persons raises a fair point. Indeed, this issue rings especially true when it comes to neuroprediction rather than actuarial prediction more generally. So, we will focus primarily on neuroprediction in this section. In addressing this issue, it is important to start by placing the importance of respecting offenders within the broader context of the other people whose interests are also at stake. After all, the offender's interests are not necessarily the only ones worth protecting; his rights to autonomy and privacy must be weighed alongside those of his potential victims. So, the relevant question is not whether the offender's rights have been violated, it is how many violations of the autonomy and privacy of the offender's potential victims equal the violation incurred by that offender. This is a difficult normative question, but it bears on legal judgments whether risk assessment is employed.

We have argued that the respect for persons is not unconditional, and the courts agree. In *Katz v. The United States* (389 U.S. 347 [1967]) the Supreme

Court ruled that in matters of public safety, some violations of privacy may be justified. For example, defendants may sometimes be required to undergo diagnostic tests and even medical surgeries in order to seize potentially probative evidence such as a bullet or ingested jewelry (Hanley, 2017). Courts have also compelled defendants to undergo psychiatric evaluations and other risk assessment measures. The courts recognize some limits to privacy, and they are not likely to hold potentially intrusive brain measures to a different standard than other types of intrusive measures (Nadelhoffer et al., 2012).

The respect for persons concern, however, presents another challenge, namely that neurotechnologies are different from other kinds of legal evidence because they reveal not just physical information but also information in the more privileged mental domain. To be sure, some neurotechnologies purport to read not just the brain but the mind. Examples of this so-called mind-reading technology include techniques for detecting lies (Schauer, 2010) or reconstructing a person's past visual experiences (Nishimoto et al., 2011). Using such techniques to draw inferences about a defendant's mental experiences has been the subject of strong scientific and ethical criticism (Ienca & Andorno, 2017), and, unsurprisingly, legal scholars have been cautious about them (Greely & Illes, 2007; cf. Schauer, 2010). Yet, even setting these criticisms aside, the question still remains of whether neurotechnologies can ever measure *mental* phenomena. If so, they might qualify as testimonial evidence (i.e., information that is evoked by questioning by the government), which might be more privileged than information that exists independently of prompting. This is important because the Fifth Amendment specifically protects defendants from being forced to testify against themselves (Ienca & Andorno, 2017; Nadelhoffer et al., 2012).

From a broad scientific perspective, neurotechnologies measure lots of different things, and many of these are better described as physical than mental properties, depending on what conclusions are to be drawn from them (Shen, 2013). The brain is responsible for myriad physical processes, such as metabolizing fat and regulating the heart, which aren't directly related to a person's subjective self. A scientist using neural markers to predict whether someone will develop Alzheimer's disease, for example, might be interested in future health outcomes without having any interest in that person's character traits or the content of her thoughts.

Appealing to a physical level of analysis, some scholars have suggested that brain information should not necessarily be privileged. People can be

sources of information about themselves without necessarily implicating themselves. This distinction allows courts to use people's personal information to some degree without violating their Fifth Amendment rights (Farahany, 2012a). *Pennsylvania v. Muniz* (496 U.S. 582 [1990]) illustrates this point. The court decided that those suspected of driving under the influence of alcohol or drugs do not need to be read their Miranda rights before their sobriety is assessed because the relevant inference concerns "the physiological functioning of [their] brain," and the brain constitutes physical information (Shen, 2013). Similarly, in *Schmerber v. California* (384 U.S. 757 [1966]), the court used information from the defendant's blood test to determine his blood-alcohol level while he was hospitalized. He was too intoxicated to consent, but the court determined that the blood-alcohol evidence was admissible because that information existed independently of his testimonial capacities and did not need his participation or effort to be determined (Fox, 2009).

When it comes to the brain, however, others have argued that the information captured by neurobiological measures may in fact include mental information because the brain directly gives rise to the mind, including a person's sense of identity and self (Ienca & Andorno, 2017). If so, it may qualify as the more privileged testimonial evidence. To qualify as testimonial evidence under current precedent, the information must describe a conscious, intentional proposition, a characteristic that is privileged in part because it is relatively easy for a defendant to conceal (Fox, 2009). Defendants, for instance, may remain silent and refuse to cooperate with lawyers and even psychologists who attempt to extract guilty knowledge for court uses. So, if brain measures could potentially reveal information about this type of information (e.g., beliefs, feelings, memories), it is easy to see why neurotechnologies might be held to a higher standard than more clearly physical measures such as breathalyzer tests.

Though the law has distinguished physical evidence from testimonial evidence, advances in neurobiology have begun to blur the boundary. Using physical measures of brain function to draw inferences about mental states would seem to raise the testimonial flag. However, not all mental information is necessarily testimonial (or incriminating). Thus, some scholars have criticized the physical–testimonial dichotomy on the grounds that it doesn't sufficiently protect citizens from potential privacy intrusions (Farahany, 2012a; Pustilnik, 2013). Farahany offers an alternative model, distinguishing

four types of brain–mind information, relayed here from least to most sacred: identifying information (using brain information to determine someone's identity, such as DNA profiling), automatic information (subconscious physiological, neurological, and reflexive patterns of behavior), memorialized information (memories determined by measures such as voice recognition tests, except those created under government coercion), and uttered information (conscious thoughts communicated through language).

Neuropredictive information, such as neural markers of behavioral traits such as impulsivity or emotional instability, is clearly not uttered. And unlike uttered information (or testimonial evidence more broadly), behavioral traits are not so easily veiled. On the other hand, such neuropredictive measures can be used to discern cognitive information beyond the subject's mere identity. This leaves automatic and memorialized information. In order to qualify as memorialized information, neuropredictive assessments would have to glean insight into the content or quality of a person's memories. But while significant strides have been made to capture neural signatures of episodic memories in the lab (e.g., Nishimoto et al., 2011; Rissman, Greely, & Wagner, 2010), in practice, neuropredictive assessments generally do not purport to probe subjective experiences. At best, they could provide indirect evidence of a behavioral trait. To qualify as memorialized information, the results of neuropredictive assessments would also need to be easily concealed. But again, while it might be possible for some individuals to refuse to cooperate or to learn to generate erroneous neural signals effortfully, the neural pattern of a typical cooperating subject would be a relatively reflexive, emergent, and cognitively impenetrable property of that person's neurocognitive functioning. Moreover, the behavioral information inferred by these tests, such as impulsivity, might already be somewhat discernible from public records, including academic, job, and criminal history, and from general interactions with others. Under this framework, neuropredictive measures that estimate risk of reoffending based on theoretical traits such as impulsivity would seem to qualify best as "automatic information." Thus, compared to some other kinds of evidence, they might be less eligible for protection. While people have a reasonable expectation of privacy of their brain function in their daily lives, if this information is classified as automatic, some of those expectations of privacy might plausibly be overridden in cases of public security.

Still, the use of neuropredictive measures to bypass people's conscious thoughts and preferences could raise new concerns, particularly if these measures are used to draw inferences about the content of the defendant's mental experiences or if they are physically intrusive (Farahany, 2012b). That said, it is not clear that functional imaging data could be collected from an unwilling offender in practice, since he could simply be noncompliant, refuse to be still in the scanner, and so on. But if such technologies could be used without the consent of the defendant (e.g., with some kind of sedatives or restraints) or without the aid of additional confirmatory/disconfirmatory evidence, great caution must be exercised to evaluate and regulate to what end the technology is to be used, since the particular conclusions that are drawn will likely vary widely from case to case as a function of the fact finder's specific question.

6.5 Conclusions and Future Directions for Risk Assessment

We began this chapter with a special interest in the ethical implications of neurobiological factors in offender risk assessments. However, we have learned that many of the concerns about such tools, such as concerns about justice and beneficence, apply to the use of offender risk assessment more generally. So, any attempt to evaluate the use of neurobiological risk assessment tools must first answer to those broader concerns. Much of the existing discourse on offender risk assessment has been focused on potential sentencing applications of actuarial risk assessment after a conviction. So, our present discussion has focused mainly on that case, where concerns about justice and beneficence seem justifiable in the abstract. But when legal actors carry an obligation to make a decision with real consequences for public safety—as is common for many risk assessment contexts—these ethical concerns apply at least as strongly to decision protocols that bypass the evidence about risk.

Despite the scholarly emphasis on sentencing applications of actuarial risk assessment, in our view, this emphasis is somewhat misdirected. The use of actuarial risk assessment for sentencing is an extreme case where a conflict of interest exists between the offender and society, delivery of a sentencing judgment is mandatory, and the stakes on both sides are high. This disproportionate emphasis in sentencing scholarship has overshadowed the variety of lower-conflict, lower-stakes legal contexts that could usefully

implement actuarial risk assessment in more obvious ways (see Baum & Savulescu, 2013). These include decisions to reduce a charge; decisions to offer bail, probation, or parole in lieu of jail or prison time; decisions to place an inmate in a lower security level; decisions about early release from civil commitment; diversion from traditional court to drug court; or the provision of benign treatment services. In all of these cases, the default choice (e.g., neglecting to offer treatment) is generally less favorable to the offender, and the offender qualifies for the more favorable choice by demonstrating low risk. Certainly, concerns could still be raised in these cases, but these concerns would likely apply to an even greater extent when such choices are not offered. Scholars increasingly agree that more attention should be paid to the uses of risk assessment that serve rather than violate the offender's interests (Baum & Savulescu, 2013; Tonry, 2014).

Neurobiological tools may be especially valuable in such contexts. Although accurate prediction does not necessarily depend on knowing the causes of target behaviors, it can often help to illuminate such causes and thus could inform the development of tailored treatments that more effectively manage those behaviors (see also Barabas et al., 2018; Douglas et al., 2017; Gilligan & Lee, 2005; Latessa et al., 2010; Meynen, 2018). Two recent brain stimulation studies demonstrate a proof of this concept. In one study using transcranial direct-current stimulation, stimulation of the ACC resulted in improved performance on cognitive and affective attentional tasks (To et al., 2018). In another study using the same technology, a single stimulation session of the dorsolateral prefrontal cortex reduced aggressive criminal intentions twenty-four hours post stimulation (Choy, Raine, & Hamilton, 2018). As with any new treatment intervention, the value of such neurobiological approaches must meet high standards of validity and reliability (Large & Nielssen, 2017).

Whether brain measures are included in risk assessments or not, actuarial risk assessment cannot be judged using a monolithic ethical rule because its uses are heterogenous, reflecting plural often competing social values such as public safety versus civil liberties. Thus, its standards for evaluation must aspire to balance these competing values, and they must do so in a manner that is responsive to the particular context of their intended use.

When evaluating the justifiability of a given use of actuarial risk assessment, it would seem helpful to receive guidance from a formal ethical theory that offers ways to balance the many competing rights and values at play in risk assessment settings. Alas, no such universally accepted theory

exists. Meanwhile, many shorter-term practical considerations remain for minimizing damage in the application of actuarial risk assessment.

First, to minimize classification errors, scientists and practitioners of actuarial risk assessment should demand large, well-matched validation samples (Douglas et al., 2017; for a review, see Poldrack et al., 2018). The larger and better matched the validation sample, the more predictive utility can be leveraged from smaller ontological categories. Relatedly, researchers should more heavily prioritize independent replication among prospective follow-up studies and randomized controlled trials whenever possible. In practice, such opportunities are often limited by funding constraints and policy regulations, but their value cannot be overstated.

Second, unlike traditional risk assessment methods, when an actuarial risk assessment classification error is detected, it is possible to track the likely source of the error so that the model can be strategically improved. This is a unique quality control advantage of actuarial risk assessment. Currently, many risk assessment tools and error rates are protected under rules of propriety. However, in criminal justice contexts where stakes are high, transparency of process is a necessary part of keeping these processes honest. Researchers and providers should be obligated to disclose all statistical and algorithmic information to practitioners and potentially the public (Hamilton, 2015; Wykstra, 2018).

Third, the use of dynamic factors in risk models should be upheld whenever possible (Wolpe, 2009) to increase self-efficacy in the offender's rehabilitative efforts while reducing stigmatization and perceptions that the offender is incorrigible.

Fourth, actuarial risk tools are only as good as their measures. Rather than ignore problematic factors such as race, developers of these tools should consider statistically controlling for such factors. Similarly, there are known problems with the use of official records as proxies for actual criminal histories. Developers should consider ways to supplement the official records of their validation samples with other measures such as anonymized self-reported criminal history.

Fifth, the criminal justice system needs clearly defined normative standards for classification accuracy. Appropriate standards will remain a moving target and will depend on a deliberate ongoing dialogue between policy makers, scientists, and practitioners. These standards must be sensitive to the different applications of actuarial risk assessment and should include

contingencies for cases in which a risk assessment result is underpowered or inconclusive (see also Campbell & Eastman, 2014).

Finally, efforts to understand our collective and often conflicting normative attitudes toward different forms of risk assessment will benefit strongly from actuarial methods that help clarify the relative costs and benefits of each alternative. If we choose to exclude the evidence from our risk assessment practices, its harms do not go away—they merely operate under the cover of darkness.

Acknowledgments

We thank Amanda Haskell, Saskia Verkiel, Nicholas Alonso, Lyn Gaudet, Jason Kerkmans, Douglas Mossman, Felipe De Brigard, Walter Sinnott-Armstrong, Natalia Washington, and the participants of the Duke Summer Seminars on Neuroscience and Philosophy. Contributions to this chapter were supported in part by a grant from the National Science Foundation: #1829495.

Notes

1. Barefoot v. Estelle, *Id*. at § 901.

2. Several recent meta-analyses have further established the predictive validity of VRAG (see, e.g., Singh, Grann, & Fazel, 2011; Singh et al., 2014; Yang, Wong, & Coid, 2010). It is worth noting, however, that VRAG was designed using a male population. Some recent research suggests its use to assess the risk of female offenders may be problematic (Hastings et al., 2011).

3. For a recent review of the literature on the biopsychosocial model of violence, see Raine (2013).

4. It is worth pointing out that actuarial tools such as VRAG do not use predictor variables such as race precisely because the models would then be held to higher standards of legal scrutiny, thereby problematizing their use in legal contexts.

References

Aharoni, E., Anderson, N. E., Barnes, J. C., Allen, C. H., & Kiehl, K. A. (2019). Mind the gap: Toward an integrative science of the brain and crime. *BioSocieties, 14*(3), 463–468.

Aharoni, E., Mallett, J., Vincent, G. M., Harenski, C. L., Calhoun, V. D., Sinnott-Armstrong, W., . . . & Kiehl, K. A. (2014). Predictive accuracy in the neuroprediction of rearrest. *Social Neuroscience, 9*(4), 332–336.

Aharoni, E., Vincent, G. M., Harenski, C. L., Calhoun, V. D., Sinnott-Armstrong, W., Gazzaniga, M. S., & Kiehl, K. A. (2013). Neuroprediction of future rearrest. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 6223–6228.

Andrews, D. A. (2006). Enhancing adherence to risk–need–responsivity: Making quality a matter of policy. *Criminology and Public Policy, 5*(3), 595–602.

Aos, S., Miller, M., & Drake, E. (2006). Evidence-based public policy options to reduce future prison construction, criminal justice costs, and crime rates. *Federal Sentencing Reporter, 19*(4), 275.

Banks, S., Robbins, P. C., Silver, E., Vesselinov, R., Steadman, H. J., Monahan, J., . . . & Roth, L. H. (2004). A multiple-models approach to violence risk assessment among people with mental disorder. *Criminal Justice and Behavior, 31*(3), 324–340.

Barabas, C., Virza, M., Dinakar, K., Ito, J., & Zittrain, J. (2018). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. *Proceedings of Machine Learning Research, 81*, 62–76.

Baum, M., & Savulescu, J. (2013). Behavioral biomarkers: What are they good for? In I. Singh, W. P. Sinnott-Armstrong, & J. Savulescu (Eds.), *Bioprediction, biomarkers, and bad behavior: Scientific, legal, and ethical challenges* (pp. 12–41). Oxford: Oxford University Press.

Bedard, H. L. (2017). The potential for bioprediction in criminal law. *The Columbia Science and Technology Law Review, XVIII*, 58.

Berns, G. S., & Moore, S. E. (2012). A neural predictor of cultural popularity. *Journal of Consumer Psychology, 22*(1), 154–160.

Berryessa, C. M. (2017). Jury-eligible public attitudes toward biological risk factors for the development of criminal behavior and implications for capital sentencing. *Criminal Justice and Behavior, 44*(8), 1073–1100.

Buckholtz, J., & Meyer-Lindenberg, A. (2012). Psychopathology and the human connectome: Toward a transdiagnostic model of risk for mental illness. *Neuron, 74*(6), 990–1004.

Camchong, J., Stenger, A., & Fein, G. (2012). Resting-state synchrony during early alcohol abstinence can predict subsequent relapse. *Cerebral Cortex, 23*(9), 2086–2099.

Campbell, C., & Eastman, N. (2014). The limits of legal use of neuroscience. In W. Sinnott-Armstrong, I. Singh, & J. Savulescu (Eds.), *Bioprediction, biomarkers, and bad behavior: Scientific, legal, and ethical challenges* (pp. 91–117). Oxford: Oxford University Press.

Cecil, K. M., Brubaker, C. J., Adler, C. M., Dietrich, K. N., Altaye, M., Egelhoff, J. C., . . . Lanphear, B. P. (2008). Decreased brain volume in adults with childhood lead exposure. *PLoS Medicine, 5*(5), e112.

Choy, O., Raine, A., & Hamilton, R. H. (2018). Stimulation of the prefrontal cortex reduces intentions to commit aggression: A randomized, double-blind, placebo-controlled, stratified, parallel-group trial. *Journal of Neuroscience, 38*(29), 6505–6512.

Coppola, F. (2018). Mapping the brain to predict antisocial behaviour: New frontiers in neurocriminology, "new" challenges for criminal justice. *UCL Journal of Law and Jurisprudence—Special Issue, 1*(1), 103–126.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674.

de Hevia, M.-D., & Spelke, E. S. (2009). Spontaneous mapping of number and space in adults and young children. *Cognition, 110*(2), 198–207.

Delfin, C., Krona, H., Andiné, P., Ryding, E., Wallinius, M., & Hofvander, B. (2019). Prediction of recidivism in a long-term follow-up of forensic psychiatric patients: Incremental effects of neuroimaging data. *PLoS One, 14*(5), e0217127.

Douglas, T., Pugh, J., Singh, I., Savulescu, J., & Fazel, S. (2017). Risk assessment tools in criminal justice and forensic psychiatry: The need for better data. *European Psychiatry: The Journal of the Association of European Psychiatrists, 42*, 134–137.

Ellwood, C. A. (1912). Lombroso's theory of crime. *Journal of the American Institute of Criminal Law and Criminology, 2*(5), 716.

Ennis, G., & Litwack, R. (1974). Psychiatry and the presumption of expertise: Flipping coins in the courtroom. *California Law Review, 62*, 693–718.

Faigman, D. L., Monahan, J., & Slobogin, C. (2013). Group to individual (G2i) inference in scientific expert testimony. *The University of Chicago Law Review, 81*(2), 417–480.

Farahany, N. A. (2012a). Incriminating thoughts. *Stanford Law Review, 64*, 351–408.

Farahany, N. A. (2012b). Searching secrets. *University of Pennsylvania Law Review, 160*, 1239.

Fox, D. (2009). The right to silence as protecting mental control. *Akron Law Review, 42*(3), 763.

Gilligan, J., & Lee, B. (2005). The Resolve to Stop the Violence Project: Reducing violence in the community through a jail-based initiative. *Journal of Public Health, 27*(2), 143–148.

Greely, H. T., & Illes, J. (2007). Neuroscience-based lie detection: The urgent need for regulation. *American Journal of Law and Medicine,* 33(2–3), 377–431.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293–323.

Hall, W. D., Carter, A., & Yücel, M. (2014). Ethical issues in the neuroprediction of addiction risk and treatment response. In J. Clausen & N. Levy (Eds.), *Handbook of neuroethics* (pp. 1025–1044). Dordrecht, Netherlands: Springer.

Hamilton, M. (2015). Risk-needs assessment: Constitutional and ethical challenges. *American Criminal Law Review, 52*, 61.

Hanley, B. J. (2017). In search of evidence against a criminal defendant: The constitutionality of judicially ordered surgery. *The Catholic Lawyer, 22*(4), 6.

Hastings, M. E., Krishnan, S., Tangney, J. P., & Stuewig, J. (2011). Predictive and incremental validity of the Violence Risk Assessment Guide scores with male and female inmates. *Psychological Assessment, 23*(1), 174–183.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: John Wiley.

Heilbrun, K., Douglas, K. S., & Yasuhara, K. (2009). Violence risk assessment: Core controversies. In J. L. Skeem, K. S. Douglas, & S. O. Lilienfeld (Eds.), *Psychological science in the courtroom: Consensus and controversy* (pp. 333–357). New York: Guilford Press.

Hubbard, R. A., Kerlikowske, K., Flowers, C. I., Yankaskas, B. C., Zhu, W., & Miglioretti, D. L. (2011). Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography. *Annals of Internal Medicine, 155*(8), 481.

Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy, 13*(1), 5.

Janes, A. C., Pizzagalli, D. A., Richardt, S., Frederick, B. de B., Chuzi, S., Pachas, G., . . . Kaufman, M. J. (2010). Brain reactivity to smoking cues prior to smoking cessation predicts ability to maintain tobacco abstinence. *Biological Psychiatry, 67*(8), 722–729.

Janus, E. S., & Prentky, R. A. (2003). Forensic use of actuarial risk assessment with sex offenders: Accuracy, admissibility and accountability. *American Criminal Law Review, 40*, 1443.

Jurjako, M., Malatesti, L., & Brazil, I. A. (2018). Some ethical considerations about the use of biomarkers for the classification of adult antisocial individuals. *International Journal of Forensic Mental Health, 18*(3), 228–242.

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D. L., Cha, C., Nock, M. K., & Brent, D. (2017). Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature Human Behaviour, 1*(12), 911.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science, 303*(5660), 1023–1026.

Kiehl, K. A., Anderson, N. E., Aharoni, E., Maurer, J. M., Harenski, K. A., Rao, V., . . . & Kosson, D. (2018). Age of gray matters: Neuroprediction of recidivism. *NeuroImage: Clinical, 19*, 813–823.

Kogut, T., & Ritov, I. (2005). The "identified victim" effect: An identified group, or just a single individual? *Journal of Behavioral Decision Making, 18*(3), 157–167.

Krauss, D. A., & Sales, B. D. (2001). The effects of clinical and scientific expert testimony on juror decision making in capital sentencing. *Psychology, Public Policy, and Law, 7*(2), 267–310.

Kunda, Z., & Nisbett, R. E. (1986). The psychometrics of everyday life. *Cognitive Psychology, 18*(2), 195–224.

Large, M., & Nielssen, O. (2017). The limitations and future of violence risk assessment. *World Psychiatry, 16*(1), 25–26.

Latessa, E. J., Lovins, L. B., & Smith, P. (2010). *Follow-up evaluation of Ohio's community based correctional facility and halfway house programs: Program characteristics supplemental report*. Cincinnati, OH: University of Cincinnati.

Lidz, C. W., Mulvey, E. P., & Gardner, W. (1993). The accuracy of predictions of violence to others. *JAMA, 269*(8), 1007.

Liu, J., Xu, X., Wu, K., Piao, Z., Huang, J., Guo, Y., . . . & Huo, X. (2011). Association between lead exposure from electronic waste recycling and child temperament alterations. *Neurotoxicology, 32*(4), 458–464.

Lombroso, C. (2006). Criminal literature. In M. Gibson & N. H. Rafter (Trans.), *Criminal man* (pp. 79–80). Durham, NC: Duke University Press.

MacKenzie, D. L. (2006). *What works in corrections: Reducing the criminal activities of offenders and delinquents*. Cambridge: Cambridge University Press.

Meynen, G. (2018). Forensic psychiatry and neurolaw: Description, developments, and debates. *International Journal of Law and Psychiatry, 65*, 101345.

Monahan, J. (1995). *The clinical prediction of violent behavior*. Lanham, MD: Jason Aronson, Inc. (Originally published in 1981 as *Predicting violent behavior: An assessment of clinical techniques* by Sage).

Monahan, J. (2014). The inclusion of biological risk factors in violence risk assessments. In I. Singh, W. P. Sinnott-Armstrong, & J. Savulescu (Eds.), *Bioprediction, biomarkers, and bad behavior: Scientific, legal, and ethical challenges* (pp. 57–76). Oxford: Oxford University Press.

Monahan, J., Brodsky, S. L., & Shan, S. A. (1981). *Predicting violent behavior: An assessment of clinical techniques*. Beverly Hills, CA: Sage.

Monahan, J., Steadman, H., Silver, E., Applebaum, P. S., Clark-Robbins, A., Mulvey, E. P., . . . Banks, S. (2001). *Rethinking risk assessment: The MacArthur Study of mental disorder and violence*. Oxford: Oxford University Press.

Monahan, J., & Silver, E. (2003). Judicial decision thresholds for violence risk management. *International Journal of Forensic Mental Health, 2*(1), 1–6.

Moore, K. E., Stuewig, J. B., & Tangney, J. P. (2016). The effect of stigma on criminal offenders' functioning: A longitudinal mediational model. *Deviant Behavior, 37*(2), 196–218.

Nadelhoffer, T., Bibas, S., Grafton, S., Kiehl, K. A., Mansfield, A., Sinnott-Armstrong, W., & Gazzaniga, M. (2012). Neuroprediction, violence, and the law: Setting the stage. *Neuroethics, 5*(1), 67–99.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology, 21*(19), 1641–1646.

Pardini, D. A., Raine, A., Erickson, K., & Loeber, R. (2014). Lower amygdala volume in men is associated with childhood aggression, early psychopathic traits, and future violence. *Biological Psychiatry, 75*(1), 73–80.

Paulus, M. P., Tapert, S. F., & Schuckit, M. A. (2005). Neural activation patterns of methamphetamine-dependent subjects during decision making predict relapse. *Archives of General Psychiatry, 62*(7), 761.

Pinker, S. (2016). *The blank slate: The modern denial of human nature*. New York: Viking.

Poldrack, R. A., Monahan, J., Imrey, P. B., Reyna, V., Raichle, M. E., Faigman, D., & Buckholtz, J. W. (2018). Predicting violent behavior: What can neuroscience add? *Trends in Cognitive Sciences, 22*(2), 111–123.

Pustilnik, A. C. (2013). Neurotechnologies at the intersection of criminal procedure and constitutional law. In J. T. Parry & L. S. Richardson (Eds.), *The constitution and the future of criminal justice in America* (pp. 109–134). Cambridge: Cambridge University Press.

Quinsey, V. L., Harris, G. E., Rice, M. E., & Cormier, C. (1998). *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association.

Raine, A. (2013). *The anatomy of violence: The biological roots of crime*. New York: Pantheon Books.

Rissman, J., Greely, H. T., & Wagner, A. D. (2010). Detecting individual memories through the neural decoding of memory states and past experience. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 9849–9854.

Ryazanov, A. A., Knutzen, J., Rickless, S. C., Christenfeld, N. J. S., & Nelkin, D. K. (2018). Intuitive probabilities and the limitation of moral imagination. *Cognitive Science, 42*(Suppl. 1), 38–68.

Sanders, T., Liu, Y., Buchner, V., & Tchounwou, P. B. (2009). Neurotoxic effects and biomarkers of lead exposure: A review. *Reviews on Environmental Health, 24*(1), 15–45.

Schauer, F. (2010). Neuroscience, lie-detection, and the law: Contrary to the prevailing view, the suitability of brain-based lie-detection for courtroom or forensic use should be determined according to legal and not scientific standards. *Trends in Cognitive Sciences, 14*(3), 101–103.

Shen, F. X. (2013). Neuroscience, mental privacy, and the law. *Harvard Journal of Law and Public Policy, 36*(2), 653–713.

Silver, E., & Miller, L. L. (2002). A cautionary note on the use of actuarial risk assessment tools for social control. *Crime and Delinquency, 48*(1), 138–161.

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., . . . & Otto, R. K. (2014). International perspective on the practical application of violence risk assessment: A global survey of 44 countries. *International Journal of Forensic Mental Health, 13*, 193–206.

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31*, 499–513.

Sinha, R., & Li, C. S. R. (2007). Imaging stress-and cue-induced drug and alcohol craving: Association with relapse and clinical implications. *Drug and Alcohol Review, 26*(1), 25–31.

Specker, J., Focquaert, F., Sterckx, S., & Schermer, M. H. N. (2018). Forensic practitioners' expectations and moral views regarding neurobiological interventions in offenders with mental disorders. *BioSocieties, 13*, 304–321.

Starr, S. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review, 66*(4), 803–872.

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Clark Robbins, P., Mulvey, E. P., . . . & Banks, S. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior, 24*(1), 83–100.

Steele, V. R., Claus, E. D., Aharoni, E., Vincent, G. M., Calhoun, V. D., & Kiehl, K. A. (2015). Multimodal imaging measures predict rearrest. *Frontiers in Human Neuroscience, 9*, 425.

Steele, V. R., Fink, B. C., Maurer, J. M., Arbabshirani, M. R., Wilber, C. H., Jaffe, A. J., . . . & Kiehl, K. A. (2014). Brain potentials measured during a go/nogo task predict completion of substance abuse treatment. *Biological Psychiatry, 76*(1), 75–83.

Steele, V. R., Maurer, J. M., Arbabshirani, M. R., Claus, E. D., Fink, B. C., Rao, V., . . . & Kiehl, K. A. (2017). Machine learning of functional magnetic resonance imaging network connectivity predicts substance abuse treatment completion. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(2), 141–149.

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British Journal of Psychology, 54*(2), 101–114.

Taxman, F. S. (2002). Supervision—Exploring the dimensions of effectiveness. *Federal Probation Journal, 66*, 14.

To, W. T., Eroh, J., Hart, J., & Vanneste, S. (2018). Exploring the effects of anodal and cathodal high definition transcranial direct current stimulation targeting the dorsal anterior cingulate cortex. *Scientific Reports, 8*(1), 4454.

Tonry, M. (2014). Remodeling American sentencing: A ten-step blueprint for moving past mass incarceration. *Criminology and Public Policy, 13*(4), 503–533.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232.

Verplaetse, J. (2009). *Localizing the moral sense: Neuroscience and the search for the cerebral seat of morality, 1800–1930*. Dordrecht, Netherlands: Springer Science & Business Media.

Weiss, B. (2017, April 1). Jonathan Haidt on the cultural roots of campus rage. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/jonathan-haidt-on-the-cultural-roots-of-campus-rage-1491000676

Wolpe, P. (2009). Is my mind mine? *Forbes*. Retrieved from https://www.forbes.com/2009/10/09/neuroimaging-neuroscience-mind-reading-opinions-contributors-paul-root-wolpe.html?sh=12d8f7936147

Wykstra, S. (2018, July). Just how transparent can a criminal justice algorithm be? *Slate Magazine*. Retrieved from https://slate.com/technology/2018/07/pennsylvania-commission-on-sentencing-is-trying-to-make-its-algorithm-transparent.html

Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136*, 740–767.

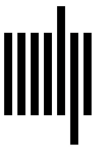
Neuroscience and Philosophy

Edited by: Felipe De Brigard, Walter Sinnott-Armstrong

Citation:

Neuroscience and Philosophy
Edited by: Felipe De Brigard, Walter Sinnott-Armstrong
DOI: 10.7551/mitpress/12611.001.0001
ISBN (electronic): 9780262367332
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif by Westchester Publishing Services. .

Library of Congress Cataloging-in-Publication Data

Names: Brigard, Felipe de, editor. | Sinnott-Armstrong, Walter, 1955– editor.
Title: Neuroscience and philosophy / edited by Felipe De Brigard and Walter Sinnott-Armstrong.
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Includes bibliographical references and index.
Identifiers: LCCN 2021000758 | ISBN 9780262045438 (paperback)
Subjects: LCSH: Cognitive neuroscience—Philosophy.
Classification: LCC QP360.5 .N4973 2022 | DDC 612.8/233—dc23
LC record available at https://lccn.loc.gov/2021000758