On the Ethics and Practicalities of Artificial Intelligence, Risk Assessment, and Race

Neil R. Hogan, PhD, Ethan Q. Davidge, MA, and Gabriela Corabian, PhD

Artificial intelligence (AI) has been put forth as a potential means of improving and expediting violence risk assessment in forensic psychiatry. Furthermore, it has been proffered as a means of mitigating bias by replacing subjective human judgements with unadulterated data-driven predictions. A recent ethics analysis of AI-informed violence risk assessment enumerated some potential benefits, ethics concerns, and recommendations for further discussion. The current review builds on this previous work by highlighting additional important practical and ethics considerations. These include extant technology for violence risk assessment, paradigmatic concerns with the application of AI to risk assessment and management, and empirical evidence of racial bias in the criminal justice system. Emphasis is given to problems of informed consent, maleficence (e.g., the known iatrogenic effects of overly punitive sanctions), and justice (particularly racial justice). AI appears well suited to certain medical applications, such as the interpretation of diagnostic images, and may well surpass human judgement in accuracy or efficiency with respect to some important tasks. Caution is necessary, however, when applying AI to processes like violence risk assessment that do not conform clearly to simple classification paradigms.

J Am Acad Psychiatry Law 49(3) online, 2021. DOI:10.29158/JAAPL.200116-20

Key words: risk assessment; artificial intelligence; ethics; race

Violence risk assessments conducted by forensic psychiatrists, as well as by other mental health and criminal justice professionals, have wide-ranging impacts on individual rights and community safety. Risk assessment information can influence decisions ranging from the relatively mundane, such as daily outings for forensic inpatients, to decisions with life and death consequences, such as capital punishment determinations. Thus, technologies that show promise for improving the process of risk assessment, whether by increasing accuracy, efficiency, or reliability, are of great interest to professionals and policy-makers alike. As argued by Cockerill[1] in a recent issue of this journal, artificial intelligence (AI) represents one such promising technology. Cockerill did a commendable job of introducing the technology and potential applications in forensic psychiatry, in addition to identifying potential ethics implications of its adoption. We felt compelled to expand on this analysis, particularly with regard to race and risk assessment. Our analysis explicates further ethics and practical implications of AI in violence risk assessment, as informed by the broader violence risk assessment and criminological literature. It is hoped that the following discussion may facilitate the critical ethics evaluation of emerging technologies in this area.

Published online June 3, 2021.

Dr. Hogan is Program Coordinator, Integrated Threat and Risk Assessment Centre, Alberta Ministry of Justice and Solicitor General, and Professional Affiliate, University of Saskatchewan. Mr. Davidge is a Certified Threat Assessor, Integrated Threat and Risk Assessment Centre, Alberta Ministry of Justice and Solicitor General. Dr. Corabian is a Clinical Forensic Psychologist, Northern Alberta Forensic Psychiatry Program, Alberta Health Services. Address correspondence to: Neil R. Hogan, PhD, Integrated Threat and Risk Assessment Centre, ALERT West Campus, T5S OC1, Edmonton, Alberta, Canada. E-mail: neil.hogan@usask.ca.

Disclosures of financial or other potential conflicts of interest: None.

Lessons Learned in Violence Risk Assessment

To evaluate the potential contributions of new technology fairly, one must first take stock of extant

technology. Cockerill asserted "reliable and accurate assessment of violence risk remains an elusive goal for forensic psychiatrists" (Ref. 1, p 345) and cited the proliferation of risk instruments as evidence. The notion that the absence of a single standard for violence risk assessment constitutes evidence of an underdeveloped field warrants further consideration. Violence is a complex concept. The definition of violence adopted by the World Health Organization[2] encompasses behaviors ranging widely in form (e.g., sexual versus nonsexual violence), motivation (e.g., instrumental versus reactive aggression), and severity (e.g., serious threats, pushing, homicide). Thus, violence constitutes many types of behavior that are complex, dynamic, and context-dependent. It follows then that the task of violence risk assessment is also nuanced and complex.

Two decades ago, Steadman[3] addressed the state of the field of violence risk assessment in this journal. He argued that the field had shifted from binary prediction paradigms to probabilistic conceptualizations of risk, and that assessments are best suited to rank-ordering individuals to match them with commensurate and graduated interventions. Steadman[3] also argued for specificity in violence risk appraisals with regard to time scale, context, and the nature of the behavior of concern. To be clear, Steadman's call for specificity did not imply that risk tools could predict discrete behaviors or events at particular dates or times; on the contrary, risk assessment tools tend to rely on relatively stable characteristics to differentiate among groups with greater and lesser proclivities toward violence. Given that discrete behaviors are overdetermined, dynamic, and context-dependent, reliable prediction is difficult. Therefore, the purpose of contemporary violence risk assessment is risk management, not prediction for its own sake.[3]

While skepticism regarding structured risk assessment undoubtedly remains in some quarters, there is a clear trend in the research literature indicating that such approaches are superior to alternatives (i.e., unstructured clinical judgment). Furthermore, research in forensic psychiatric settings indicates that, in keeping with Steadman's sage advice, when specialized tools are used to address specific outcomes (e.g., imminent[4] versus longer-term inpatient aggression[5] versus community recidivism[6]), these instruments can discriminate reliably between higher- and lower-risk individuals. In this sense, the proliferation of tools in violence risk assessment parallels psychiatric services more broadly. Surely most practitioners would concur that the wide range of psychotropic and psychological interventions available to treat mental disorders simply reflects the range of etiologies, forms, and severity of those disorders. In a similar way, the range of available violence risk assessment tools, at least to some degree, reflects the diverse constellation of behaviors deemed to constitute violence. Thus, we disagree with Cockerill's early premise to some extent, in that we believe that modern risk assessment technology is sufficiently established to conclude, at a minimum, that validated tools consistently outperform unstructured clinical judgements. It is with this basic premise and understanding that we endeavored to conduct our own analysis of the ethics implications of AI-driven violence risk assessment.

AI and Existing Risk Technology

As described by Cockerill, AI refers to computer algorithms that perform functions heretofore limited to "human intelligence" (Ref. 1, p 345), while deep learning refers to a form of AI utilizing artificial neural networks to derive patterns from large datasets in an atheoretical manner. Over time, such systems essentially refine themselves, based on binary feedback regarding correct or incorrect classifications.[7] Applied to violence risk assessment, AI consists of sophisticated statistical prediction models that can combine countless data points in complex ways to identify persons at risk of violence. With sufficient data and computing power, AI could develop models with previously unfathomable complexity. Conceptually, the process of AI-driven violence risk assessment represents an enhanced version of a well-established paradigm in forensic mental health, the empirical actuarial approach (i.e., using group data regarding known outcomes to evaluate the likelihood of future outcomes). Thus, AI carries the potential to amplify or mitigate both the strengths and weaknesses associated with existing actuarial prediction techniques. Therefore, it would be prudent for proponents to be mindful of the lessons of the past and to consider how they might apply in the future.

Cockerill offered illustrative examples of both hypothetical and actual applications of AI and deep learning to classification tasks. For instance, deep learning may be used to identify a particular class of organism in photographs or to diagnose pathology in chest x-rays. As an example of the latter, an existing

deep learning algorithm receiving only binary feedback (i.e., correct or incorrect) was able to achieve a level of diagnostic accuracy comparable with that of experienced radiologists while achieving a superior degree of efficiency.[8] The prospect of expedited violence risk assessments suggested by these examples is undoubtedly appealing to many practitioners in forensic psychiatry. Whether these examples are truly analogous to the process of violence risk assessment is a matter for debate.

Deep learning systems follow an iterative process, whereby correct predictions reinforce the models and incorrect predictions prompt the models to recalibrate. This approach is well suited to the interpretation of chest x-rays or the identification of lizards because both the true positive states and the true negative states are clear. Whether a physiological abnormality was truly present (or absent) is verifiable with a high degree of accuracy. Crucially, the processes used to obtain and interpret x-ray images do not influence the true state. Similarly, photographing and identifying an organism, such as a lizard, does not alter the organism's true species. Furthermore, many physical pathologies are relatively stable, while an organism's species classification does not change. In violence risk assessment however, defining the true positive state poses a conundrum because various factors, including predictions themselves, may influence outcomes and confound predictive models. As pointed out by Norko and Baranoski, the principles of signal detection theory that underlie binary prediction paradigms are conceptually problematic because a violence risk assessment should, by design, facilitate "the process of converting the high-risk true positive into a low-risk and hopefully false positive" (Ref. 9, p 86).

To illustrate this problem, consider the case of John. A computer algorithm flagged John as posing an imminent risk for a violent act on the basis of recent disturbance complaints made to police by his neighbors, his medical records, his online activities, and various other risk factors. This flag triggered police to respond to John's home with the aim of preventing violence. Upon locating John, police officers describe his behavior as agitated and aggressive. The extent to which the presence of police influenced or escalated John's level of agitation is unclear, but it is certainly possible that such interventions and confrontations could precipitate or exacerbate the very behaviors the algorithm was created to predict. During the course of their subsequent investigation, the officers discover that John possesses a firearm and ammunition, and a review of his Internet browsing history identifies various news articles and forums focused on school shootings. The officers ultimately detain John and transport him to a forensic psychiatric hospital for further evaluation.

This example may appear to represent an illustration of the potential benefits of AI-informed violence risk assessment, given the ultimate goal of preventing violence from occurring. This example also illustrates an inherent limitation of the adaptive mechanisms of deep learning as applied to violence risk assessment and management, and a fundamental limitation of the binary prediction paradigm in violence risk assessment more broadly. Even if we take for granted the premise that it was inevitable that John was going to perpetrate an act of violence (which, practically speaking, is hard to justify), this scenario poses a problem. By virtue of the fact that John never engaged in an act of violence, his case seems to warrant classification as a false prediction by the algorithm. The algorithm would then recalibrate, thereby reducing the weighting of those factors that identified John as a high risk for violence in future predictions. Note also that while it may be tempting to code this incident as a true positive, this would undermine the process altogether. The intervention by police ensured that the predicted violence did not occur, but this outcome also ensured that it is not possible to determine whether violence would have occurred without intervention. To support the accuracy of a prediction based solely on confidence in that prediction would constitute circular and indefensible reasoning.

While the hypothetical case of John is perhaps far removed from the daily clinical activities of most forensic psychiatrists, the essence of the dilemma that the case illustrates is applicable to many contexts. If an adaptive algorithm successfully informs interventions that prevent violent acts, whether through involuntary detention or medication, increased severity in sentencing, or revocation of patients' privileges or freedoms, the interventions themselves will render impossible the very behaviors required to guide further calibration. This conundrum renders the benefits of an AI system moot, given the inability to refine the system's predictive model by means of classifying the assessment as correct or incorrect. Thus, adaptive algorithms used to predict and

ultimately prevent violence can create a negative feedback loop, whereby increasing precision that leads to success in reducing violence via timely interventions will erode the accuracy of future models. Thus, it is decidedly unlike species identification or disease classification.

Further, the absence of violence among individuals who have been incapacitated (e.g., low community recidivism among involuntarily committed patients) cannot in and of itself provide evidence that the decisions to incapacitate were correct or defensible. Consider for example the high-risk Baxstrom cohort, who were detained based on professional opinions of their dangerousness; however, after a Supreme Court decision prompted their transfer out of secure facilities, the patients demonstrated a low rate of violent recidivism.[10] Again, interventions ultimately build a black box around the accuracy of violence predictions, posing a problem for adaptive algorithms requiring clear and continuous feedback.

Actions taken to prevent crime are not the only concerns with Cockerill's proposed paradigm for AI-driven risk assessment. It is important to take a step back to consider the data that we feed into the algorithms. Even data pertaining to what are ostensibly clear outcomes, such as formal criminal justice responses to violent crimes (e.g., arrests, charges, or convictions) may lead to problematic and misleading outcome variables. Cockerill briefly introduced the question of race as it pertains to risk assessment algorithms, rightly noting that some algorithms appear to contribute to disproportionate punishments for minority group members, and that the potentially pernicious role of race in the criminal justice system is not well understood. We believe, however, that there is already sufficient empirical evidence to raise ethics concerns regarding the application of AI to risk assessment among diverse populations, and that a more detailed analysis in this regard is warranted.

There is no doubt that racial disparities exist within the criminal justice system. For instance, empirical evidence suggests that Canadian Aboriginal persons are over-represented within the nation's total offender population, more likely to spend their sentences in custody, over-represented in maximum-security institutional settings and in segregation, and experience greater rates of parole revocation.[11] Similar racial disparities are evident in the United States, such as the considerably higher incarceration rates observed among racial minority groups relative to white U.S. residents.[12] To answer the question of whether these discrepancies pose a functional or ethics problem for AI-driven risk assessment, or risk assessment in general, further information is required.

As a starting point, we acknowledge that evidence of disparity is not necessarily evidence of unfairness. If two groups perpetrate violence at different rates, and if risk assessment algorithms predict violence with similar accuracy among the two groups, then observed differences in risk scores and predictions may be justified. Given a disproportionate rate of violent offending, a disproportionate rate of punishment and intervention among a particular group also appears justified. Furthermore, it is important to acknowledge that it is not possible to infer causation from correlational data; for instance, criminal justice disparities attributed to race could be better explained by differences in socioeconomic status, or vice versa. Various authors have offered variations of these arguments in response to criticisms of racial bias in risk assessment, and some have provided empirical support. For instance, Skeem and Lowenkamp[13] reported that, while scores on an actuarial tool varied between racial groups, the meaning of those scores was essentially the same across the groups; i.e., the relationships among scores and recidivism rates, defined as rearrests, were generally consistent.

Implicit in the preceding argument is the premise that outcome variables, such as convictions or arrests for violent crimes, are themselves objective and fair metrics. Although the empirical data present a nuanced picture in this regard, a blanket endorsement of this assumption is no longer defensible. In the case of convictions, Devine and Caughlin[14] conducted a meta-analysis focused on jury decision-making and found evidence of a small racial bias effect regarding guilty verdicts. With more extensive analysis of the data, these authors observed larger bias effects when focusing on black mock jurors and white defendants, and on white mock jurors with Hispanic defendants. This observation and similar findings raise doubt about the objectivity and precision of convictions as a reflection of behavior.

For their part, Skeem and Lowenkamp[13] relied heavily on a violent arrest criterion rather than convictions in an attempt to address such concerns in their recidivism analyses, calling it "the most unbiased criterion available" (Ref. 13, p 690). A

meta-analysis of police arrest decisions conducted by Rinehart-Kochel and colleagues[15] suggested, however, that arrests also constitute a potentially problematic metric. Their data revealed that minority groups were more likely to be arrested, even after controlling for the seriousness of the alleged offense, the amount of evidence available, the presence of witnesses, and the prior record of the suspect. Even small problems with input data can pose large problems for development of AI-driven algorithms, given their intentionally atheoretical efforts to identify patterns within information. While it is true that a computer or algorithm is itself a neutral system, professionals remain ethically accountable when they rely on a system that predicts a biased criterion measure.

Analysis

For the sake of consistency, we will follow Cockerill's lead and revisit the ethics principles of autonomy, beneficence and nonmaleficence, and justice put forth by Beauchamp and Childress.[16]

Autonomy

Cockerill's review raised a number of key questions in this domain, such as the use of personal information, including personal health information, to inform AI-driven risk assessment tools. A hypothetical case example was provided in which a young man ("Kyle") ostensibly developed an interest in cannibalism after watching the film "The Silence of the Lambs." The case was brought to the attention of a physician after being flagged by a risk assessment algorithm that analyzed publicly available information; after an involuntary hospitalization, Kyle described the physician's decision to detain him as "a terrible injustice." In discussing the case, Cockerill emphasized the tension between an individual's presumed right to privacy with regard to certain matters (e.g., Internet browsing history), and the responsibility of health professionals to protect the patient and others from potential harm. As noted, HIPAA provides for violations of patients' right to privacy in certain circumstances based on threats deemed "serious and imminent," but individual patients may dispute the legitimacy of professionals' decisions in this regard.

We concur with Cockerill's concern that, if AI-driven violence risk assessment expands into the realm of prevention among the public, concerns regarding privacy and autonomy will grow. It also bears mentioning that the ethics and jurisprudence principles that regulate health care professionals differ from those that regulate other professionals and institutions, such as law enforcement agencies and correctional institutions. Indeed, there are critical differences between a patient's right to autonomy in relation to law enforcement and public safety and a patient's right to autonomy with regard to personal care and health information. It follows then that, for the purposes of an ethics analysis, a distinction should be made between the use of AI-driven risk assessment by law enforcement agencies and the use of these technologies by health care professionals themselves. In the example provided by Cockerill, the personal information entered into the AI-driven risk assessment algorithm occurred prior to the involvement of a health care professional and did not necessarily include health information. From the physician's perspective then, this scenario does not differ in many respects from those that are already commonplace, and so typical ethics considerations related to autonomy presumptively apply. Even absent AI, the decision to detain and transport an individual to the care of a physician for assessment can be made by a police officer and may be influenced by any number of legitimate or illegitimate factors (e.g., administrative policy, "data-driven" policing, public complaints, the officer's intuition or experience). The subsequent decision of whether or not to admit the patient involuntarily for additional assessment or treatment services is separate and is the physician's own. While we agree that awareness of the police's rationale would be valuable, we also point out that most professionals are duty-bound to conduct their own assessment.

In our view, respect for the autonomy of patients in forensic psychiatry is challenged most by AI-driven risk assessment in those instances in which individuals hold a more obvious presumptive right to informed consent. Unlike public safety measures and processes that utilize data in the public domain (including police surveillance or observation of public behavior), many critical activities in forensic mental health involve an individual's direct and voluntary participation. For instance, an individual's voluntary participation in a forensic mental health evaluation typically involves an interview and, depending on the circumstances, may involve the

granting of access to protected health information, correctional records, and other collateral information sources. At present, an evaluator can provide the individual with a reasonable overview of the risks and benefits of common risk assessment procedures (e.g., the potential for the individual to receive helpful interventions versus the risk of a heightened criminal justice response) with a relatively straightforward explanation of common violence risk factors and information and tools used to ascertain their presence or absence. To the extent that AI limits evaluators' ability to comprehend the nature of their own assessments (e.g., determining which elements of the health record are being considered, and why), it also undermines their ability to explain the process to the persons being evaluated. These questions pose a significant threat to informed consent or assent.

Beneficence and Nonmaleficence

Professional practice in forensic mental health is rife with ethics dilemmas.[17] For instance, while in many contexts physicians' primary duty of care to a particular patient is clear, in forensic psychiatry the identification of the primary client is often difficult, and balancing the needs of individuals against those of the courts or society can be challenging. Thus, making decisions based on ethics is often a complex task, which involves weighing the costs and benefits of a service in relation to the individual, and in relation to communities and institutions. This problem is not new, nor is it unique to AI, but hypothetical applications of deep learning to violence risk assessment as described by Cockerill could certainly amplify or add new dimensions to the problem.

Perhaps the most obvious threat to the principle of nonmaleficence posed by AI-driven risk assessment involves situations in which use of force is guided by predictions of violence, such as when physicians alert law enforcement agencies to persons perceived to pose imminent threats of harm to others. While it is true that police interventions carry the potential to prevent or mitigate certain acts of violence, it is worth acknowledging that confrontations with law enforcement professionals also carry the potential to precipitate or exacerbate risky situations, and to elevate the immediate risk of violence in some cases. Depending on the circumstances, the unintended consequences of police use of force could contribute to immediate harm inflicted upon the subject of the assessment, the originally identified potential victims, bystanders, or the officers themselves. Notably, the potential for confrontation to escalate conflict is not limited to law enforcement and may be observed in other circumstances, such as those in which health care professionals resort to physical restraint to manage a threat of inpatient violence. While these concerns are not limited to AI-driven risk assessment, this technology carries the potential to accelerate and proliferate such preventive actions, while simultaneously obscuring the decision-making processes that guide them. Critically, if an AI-driven prediction itself contributes, directly or indirectly, to a violent incident, it creates a positive feedback loop, which further reinforces the model.

Returning to the hypothetical example provided by Cockerill, it can be argued that a likely act of imminent violence was averted in the short-term due to a deep learning algorithm. Absent such an algorithm, this individual would not have come to the attention of law enforcement or the physician who admitted him, and he may have gone on to perpetrate an abhorrent act of violence. Taking the example further though, we must consider what becomes of a person such as Kyle after the 30-day hold expires and he returns to the community. In some circumstances, such as situations in which a person's problems relate to first-episode psychosis, the outcome could be an early and effective medical intervention that benefits all involved. In other circumstances, the intervention could apply a brief pause to a long-term problem because the risk factors motivating a future act of violence remain. In such cases, the short-term intervention may accomplish little else beyond undermining that person's trust in professionals and likely responsivity to future interventions. In other cases, such as those in which persons are considering acts of violence out of disdain for governments or other institutions, such interventions could push them closer to, not further away from, translating their violent thoughts into actions. These concerns are supported by an abundance of data pertaining to the risk-need-responsivity (RNR) model of offender rehabilitation. Simply identifying those at risk for violence, regardless of the accuracy of assessments, leads to minimal impacts on overall rates of offending.[18] In contrast, meta-analytic reviews informed by the RNR model indicate that, when assessments drive interventions toward conceptually meaningful criminogenic needs

that are relevant to the individual's case, substantive reductions in offending can be achieved.[18–20]

Understandably, when faced with uncertainty regarding future violence, many clinicians take the view that it is best to err on the side of caution. In practice, this can result in a tendency to offer more rather than less intervention. Unfortunately, lessons learned from the correctional rehabilitation literature[18] also indicate that more intervention (e.g., longer periods of institutionalization) actually can increase offending rates over the long-term. Furthermore, exposing lower-risk individuals (as assessed with a structured and comprehensive risk or needs instrument) to higher-risk and psychopathic offenders tends to increase the offending rates of the former.[21,22] Such phenomena pose a conundrum for those hoping to apply AI to the general population to prevent rare events and acts of violence, particularly when the options for intervention are limited (e.g., detention in a forensic psychiatric or correctional institution). To be clear, this body of literature suggests that the potential harm caused by false positives and unnecessary interventions (e.g., arrests as opposed to warnings, or longer as opposed to shorter periods of detention[18]) may go well beyond a temporary inconvenience to individuals and actually may precipitate increased future offending rates.[18,21,22] To the extent that AI extends the reach of violence-prevention efforts into the general population, the iatrogenic effects of certain interventions will warrant consideration.

Justice

In keeping with ethics principles of justice and fairness, true objectivity has long been among the loftiest aspirations of purely actuarial risk assessment methods. Unfortunately, many of the concerns identified earlier in this article, including the potential for racial disparities to corrupt outcome measures, are pernicious problems. This problem is not unique to AI, but it could certainly be exacerbated by AI. As mentioned earlier, Cockerill and others have touched on this matter by acknowledging that algorithms are only as sound as the data they analyze. We believe that this caveat warrants explication, lest evaluators overlook its significance.

The objectivity and neutrality of data and mathematics constitute convincing support for the fairness of purely actuarial risk assessment, particularly when compared with unstructured human judgements. Challenges to these approaches based on mathematics should carry as much weight as favorable arguments, and such challenges do exist. For instance, according to Chouldechova,[23] even when other psychometric "fairness criteria" may be met, such as predictive parity (e.g., high-risk individuals reoffend at similar rates, regardless of group membership), different base rates between groups can lead to disparate impact via disproportionate error rates. Put another way, Chouldechova offered mathematical evidence indicating that, if a tool achieves predictive parity and two groups reoffend at different rates, then the rates of false positives and false negatives cannot be equal between the two groups. Thus, relying on the results of an atheoretical deep learning algorithm, with the justification that high-risk offenders of any race reoffend at similarly high rates, also requires acceptance of the reality that persons of particular racial groups will more likely be subjected to a false positive prediction, and thereby subsequent unjustified interventions, than persons of other racial groups.

To be clear, a deliberation based on ethics may lead to the conclusion that AI-driven risk assessment is more just than alternative approaches to risk assessment, particularly when the alternative under consideration is unstructured and flawed human judgment. Even the aforementioned problem with error rates is not necessarily fatal, given that it would also apply to groups with disparate recidivism rates driven by legitimate criminogenic factors, such as gang members and non-gang members. Our recommendation is simply that users and proponents of actuarial risk assessment (in which we include ourselves) give due consideration to legitimate challenges to the justice and fairness of actuarial techniques (including AI-driven techniques). Due consideration of ethics includes exploring options to mitigate and address any legitimate limitations.

We further suggest that, in light of the aforementioned racial disparities in relevant variables (e.g., arrests and convictions) that are beyond the control of risk assessors, assessors select the variables within their control carefully. Practically speaking, wherever possible, input data for any actuarial forms of risk assessment should comprise psychologically and theoretically meaningful constructs. Allowing AI systems to make atheoretical distinctions among individuals without careful consideration of the decision parameters makes the AI systems susceptible to reflecting implicit and explicit racial biases. For

instance, it is plausible that implicit biases could be reflected in criminal justice professionals' notes with regard to their perceptions of an offender's prospects for treatment, attitudes, or mental health, while also affecting decisions that influence outcomes directly and indirectly (e.g., access to community transition or employment programming). If outcome variables are in any way unfairly influenced by race, the potential exists for an AI-driven algorithm to select atheoretical and arbitrary correlates or predictors of race and present them as justifications for unfair treatment.

Returning to the analogy of chest x-rays and AI, it is reasonable to assume that if two persons with identical physical abnormalities are examined, their race should bear no significance to the likelihood of identifying said abnormality. In contrast, based on the data reviewed above, it is not possible to be certain that two persons engaging in identical violent behavior would be equally likely to register an arrest or a conviction for that behavior, even if the circumstances differed only on the perpetrators' race. Simply put, this is not justice.

In contrast, if an AI-driven algorithm utilizes data based on known and psychologically meaningful risk factors, such as attitudes that condone violence or expressed intent to perpetrate violence, then decisions to intervene are more defensible. Recent advances in the broader literature on violence risk assessment show the promise of such approaches, described as third- and fourth-generation risk tools.[24] For instance, using relatively sophisticated regression models to combine actuarial risk estimates with theoretically informed measures of cognitive and behavioral change, researchers have demonstrated that measurable treatment-related changes (and failures to change) can improve predictions of violent recidivism among forensic patients.[6] Where statistical discrepancies are unavoidable, the use of credible and meaningful risk factors mitigates threats to justice and fairness.

Conclusions

Increasingly, professionals in various fields are considering AI as a means of improving decision-making. Given the wide-ranging impacts of violence risk assessment, and the inherent complexity of the task, enthusiasm regarding the potential of AI in this domain is understandable. We agree, however, with Cockerill's conclusion that ethics considerations must be at the forefront of any efforts to apply such technology to risk assessment and management. AI seems ideally suited to certain applications, such as the interpretation of x-rays, where it could surpass alternatives based on human judgment precisely because it is unencumbered by human limitations. Caution is necessary, however, when applying AI to processes that do not seem to fit such a simple classification paradigm. For instance, as detailed above, with regard to how predictions can influence violent outcomes, they can themselves create positive and negative feedback loops that skew predictive models away from their intended aims. Additionally, as illustrated by the preceding discussion of challenges related to racial disparity, questions regarding the neutrality and objectivity of AI-driven risk assessment cannot be dismissed easily. Based on the concerns raised above, we believe that allowing an AI-driven process to operate independently at this time would risk masking and amplifying unresolved problems in atheoretical actuarial risk assessment, and this is problematic from an ethics perspective, if not indefensible.

On the other hand, with prudent and judicious oversight, AI-driven actuarial risk assessment certainly has the potential to improve upon various existing practices. Indeed, many of the problems with AI-driven risk assessment are reflections of problems that are applicable to unstructured clinical judgments. Our primary aim is to caution forensic mental health professionals against abdicating their professional and ethics responsibilities to a purportedly neutral algorithm. Toward the goal of maximizing the benefits of new technology in risk assessment, while minimizing the potential limitations, we offer the following recommendations.

Foundational training in the history of violence risk assessment (e.g., problems with binary prediction paradigms), as well as advanced training in the current state of the field, should precede any training in computer science for forensic mental health professionals conducting violence risk assessments.

Forensic mental health clinicians and researchers should take note of the broader literature on race in the criminal justice system, such as evidence for unjustified disparities in outcomes that previously may have been presumed to reflect objective metrics.

We acknowledge that an assessment comprising a comprehensive review of psychologically meaningful risk constructs is not possible in all circumstances. We further acknowledge that screening or triage

measures based on atheoretical actuarial data may substantially improve some decision-making processes. Therefore, we recommend that evaluators employ a guideline based on the best available evidence when selecting a risk assessment procedure, with the understanding that assessments with greater implications (e.g., those that inform capital punishment decisions) require a higher standard of evidence. Structured tools are considered preferable to unstructured clinical judgements. Structured tools based on both theoretically and empirically relevant risk factors are preferable to those based on empirical data alone (assuming comparable statistical properties).

Building on the previous recommendation, we suggest that forensic evaluators employ measures that may inform risk management procedures and interventions, wherever possible.

Insofar as they are in a position to observe credible evidence of discrepancies in violence risk and violent outcomes based on race or other diversity factors, mental health professionals and researchers should consider whether they have an ethics responsibility to explore and comment on the causes of such discrepancies. We share the view recently expressed by Martinez and Candilis[25] that, as forensic professionals, we can no longer purport to engage in fair and objective practices based on statistics or technology, without reckoning with societal context. By dispassionately accepting observed differences among groups that are potentially influenced by injustice as an immutable reality, we may well be shirking some of our most fundamental ethics responsibilities.

References

1. Cockerill RG: Ethics implications of the use of artificial intelligence in violence risk assessment. J Am Acad Psychiatry Law 48:345–9, 2020
2. World Health Organization: World Report on Violence and Health: Summary. 2002. Available at: https://www.who.int/violence_injury_prevention/violence/world_report/en. Accessed July 16, 2020
3. Steadman HJ: From dangerousness to risk assessment of community violence: taking stock at the turn of the century. J Am Acad Psychiatry Law 28:265–71, 2000
4. Hogan NR, Olver ME: A prospective examination of the predictive validity of five structured instruments for inpatient violence in a secure forensic hospital. Int J Forensic Ment Health 17:122–32, 2018
5. Hogan NR, Olver ME: Assessing risk for aggression in forensic psychiatric inpatients: an examination of five measures. Law Hum & Behav 40:233–43, 2016
6. Hogan NR, Olver ME: Static and dynamic assessment of violence risk among discharged forensic patients. Crim Just & Behav 46:923–38, 2019
7. Najafabadi MM, Villanustre F, Khoshgoftaar TM, *et al*: Deep learning applications and challenges in big data analytics. J Big Data 2:1, 2015
8. Rajpurkar P, Irvin J, Ball RL, *et al*: Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15:e1002686, 2018
9. Norko MA, Baranoski MV: The prediction of violence; detection of dangerousness. Brief Treat Crisis Interv 8:73–91, 2008
10. Baxstrom v Herold, 383 U.S. 107, 1966
11. Public Safety Canada: Corrections and Conditional Release Statistical Overview. Ottawa, Canada. 2015. Available at: http://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ccrso-2014/index-en.aspx. Accessed July 16, 2020
12. Bronson J, Carson EA: Prisoners in 2017: Bulletin – NCJ 252156. US Department of Justice, 2019. Available at: https://www.bjs.gov/content/pub/pdf/p17.pdf. Accessed September 23, 2020
13. Skeem JL, Lowenkamp CT: Risk, race, and recidivism: predictive bias and disparate impact. Criminology 54:680–712, 2016
14. Devine DJ, Caughlin DE: Do they matter? A meta-analytic investigation of individual characteristics and guilt judgments. Psychol Pub Pol'y & L 20:109–34, 2014
15. Rinehart-Kochel T, Wilson DB, Mastrofski SD: Effect of suspect race on officers' arrest decisions. Criminology 49:473–512, 2011
16. Beauchamp TL, Childress JF: Principles of Biomedical Ethics, Seventh Edition. New York: Oxford University Press, 2013
17. Haag AM: Ethical dilemmas faced by correctional psychologists in Canada. Crim Just & Behav 33:93–109, 2006
18. Bonta J, Andrews DA: The Psychology of Criminal Conduct, Sixth Edition. New York: Routledge, 2017
19. Dowden C, Andrews DA: Effective correctional treatment and violent reoffending: a meta-analysis. Canadian J Crim 42:449–67, 2000
20. Hanson RK, Bourgon G, Helmus L, *et al*: The principles of effective correctional treatment also apply to sexual offenders: a meta-analysis. Crim Just & Behav 36:865–91, 2009
21. Bonta J, Wallace-Capretta S, Rooney J: A quasi-experimental evaluation of an intensive rehabilitation supervision program. Crim Just & Behav 27:312–29, 2000
22. Lovins B, Lowenkamp CT, Latessa EJ: Applying the risk principle to sex offenders: can treatment make some sex offenders worse? Prison J 89:344–57, 2009
23. Chouldechova A: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. J Big Data 5:153–63, 2017
24. Bonta J, Andrews DA: Risk-Need-Responsivity Model for Offender Assessment and Rehabilitation (Report no. 2007–06). Ottawa, Canada: Department of Public Safety and Emergency Preparedness Canada, 2007. Available at: https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/rsk-nd-rspnsvty/index-en.aspx. Accessed September 23, 2020
25. Martinez R, Candilis P: Ethics in the time of injustice. J Am Acad Psychiatry Law 48:428–430, 2020