Ethics, Artificial Intelligence, and Risk Assessment

Benjamin L. Spivak, PhD, and Stephane M. Shepherd, PhD

A considerable number of papers have been published on the ethics of artificial intelligence for the purposes of violence risk assessment. In this issue of The Journal, Hogan and colleagues argue that artificial intelligence introduces novel concerns for violence risk assessment that require consideration. While the concerns that have been raised are entirely valid and require consideration, we argue that artificial intelligence does not herald a more serious or unique challenge in these areas relative to other forms of violence risk assessment.

J Am Acad Psychiatry Law 49(3) online, 2021. DOI:10.29158/JAAPL.210066-21

Key words: artificial intelligence; risk assessment; prediction; crime; recidivism; violence

Hogan *et al.*[1] have carefully identified several areas of concern with respect to the use of artificial intelligence (AI) for the purposes of assessing risk of future violence. These areas, broadly, are difficulties determining the extent to which decisions made on the basis of a risk assessment invalidate predictions; the problem of biased or noisy data contributing to biased predictions; the potential to exacerbate racial disparities within the criminal justice system; and a lack of transparency obscuring the processes utilized to make predictions. Each of these areas is tied to ethics concerns related to autonomy, beneficence and nonmaleficence, and justice. We agree with the authors that each of these areas is worthy of consideration and represents challenges for the development and application of risk assessments. Yet there is simply no convincing argument offered that AI heralds a more serious or unique challenge in any of these areas compared with other forms of risk assessment.

It is worth beginning by outlining where AI is similar to traditional approaches to risk assessment and where it differs. Artificial intelligence is a term that has numerous definitions.[2,3] For consistency with Hogan *et al.*, we will use the definition provided by Cockerill: an algorithm capable of "performing some function previously thought to be exclusive to human intelligence" (Ref. 4, p 345). This broad definition encompasses a range of algorithms utilized for a variety of functions ranging from prediction to classification to causal inference.[5] We interpret Hogan *et al.*[1] to be primarily concerned with AI applied to the task of predicting offending. Understood in this limited way, it is clear that AI may simply represent a new term for what has been established practice for decades. For example, so-called second-generation risk assessments were characterized by the use of algorithms in the form of weighted additive models to produce risk estimates.[6]

While second-generation risk assessments relied primarily on regression and actuarial tables to create their risk models, there exists a broader range of algorithms that may be useful approaches to improving AI-based prediction. For example, methods that may be broadly classed under the umbrella of supervised learning, such as random forests, artificial neural networks, stochastic gradient boosting, and support vector machines, offer alternatives to parametric regression for developing AI-based approaches to risk assessment. Another form of AI that is less commonly used in policy making settings is reinforcement learning.[7] Reinforcement learning uses positive and negative states to determine an optimal set of actions to maximize a reward (e.g., correctly classifying people who are violent). To our knowledge, this approach has never been utilized for violence risk assessment.[7] These two broad classes of algorithms

Published online 30 July, 2021.

Dr. Spivak is Lecturer and Dr. Shepherd is Associate Professor, Centre for Forensic Behavioural Science, Swinburne University of Technology, Melbourne, Victoria, Australia. Address correspondence to: Stephane Shepherd, PhD. E-mail: sshepherd@swin.edu.au.

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

appear to be generating much of the enthusiasm and much of the concern about the application of AI to risk assessment.

Where algorithms are utilized to predict an outcome, it is considered good practice to examine routinely their predictive performance on new data, a process sometimes termed validation or recalibration. This process can be automated through another algorithm that essentially runs the predictive algorithm again and determines the extent to which predictive performance has changed and the extent to which changes in the algorithm (e.g., updated beta weights for risk factors in a regression model) might better fit the new data. Automating the process of predictive validation is a task in which analysts are generally advised to proceed with caution, particularly in cases where the outcome-generating process is dynamic, as is the case with criminal offending. It is clear that, without human judgment, an entirely automated process may lead to suboptimal performance and may incorporate information in ways that have problematic ethics implications.

While automated validation can conceivably present problems if used unwisely, the same is true of manual revalidation that does not provide careful consideration to the way that data relate to offending and the extent to which other factors might affect an algorithm's predictive utility. Hogan *et al.* raise the concept of feedback loops that may occur where a risk-assessment algorithm is deployed and decisions made on the basis of algorithmic predictions shift an individual's risk level (e.g., through incapacitation or through the provision of intensive therapy).

It is worth emphasizing the obvious point that risk assessments are typically developed on populations that are not receiving an intervention on the basis of the assessment. Following deployment of a risk assessment, it may be that successful intervention on the basis of the assessment results in a reduction in offending among those classified as high risk. In this circumstance, recalibration of the algorithm without incorporating the effect of decisions made on the basis of assessment could result in individuals who would have been high risk without an intervention being labeled as lower risk. This could in turn result in these individuals not receiving the very intervention that has resulted in their being labeled as lower risk in the first place.

The problem of feedback loops exists regardless of whether one makes predictions on the basis of AI or some other method, but the problem can be circumvented to some degree through careful evaluation. Studies that incorporate a prospective design that compares those who receive an intervention on the basis of a risk assessment and those who are assessed but do not receive an intervention can provide evidence to determine the extent to which interventions premised on the result of an assessment are likely to reduce or increase the risk of individuals receiving them.[8] This information can then be incorporated in future decision-making about risk. The point that Hogan *et al.* make about feedback loops underscores the need to take into account the causal effect of decision-making that is premised on the use of risk assessment and to refrain from naively validating risk assessments that are in use without taking into account the causal effects of risk classifications.

The next major concern raised by Hogan *et al.* relates to the extent to which AI-driven risk assessment may contribute to racial disparities in the criminal justice system. The concern rests on the use of biased outcome criteria. The argument is that if an algorithm is predicting a biased outcome such as arrest, the algorithm could potentially reinforce racial bias by classifying those likely to be arrested as high risk. The concern is valid, in that the use of a biased criterion could indeed lead to concerning feedback loops that may reinforce disparities. The important question is whether the problem is with AI or with the criterion variable. It seems to us that criminal justice outcomes are generally going to be the criterion upon which any risk-assessment system, AI or otherwise, is evaluated. Therefore, concerns about biased outcomes reinforcing biases is a concern that is equally troubling regardless of how one chooses to assess risk.

As Hogan *et al.* note, recent work on fairness in risk assessment proves that, except in highly stylized situations (e.g., equal base rates between groups), it is not possible to achieve total equality between groups.[9,10] For example, if those employing risk assessment wish to have equal prediction accuracy across groups, then they must have imbalance in false positive and false negative rates. This result, however, holds regardless of whether the risk classification is made through AI, structured professional guidelines, or any other form of assessment. While total equality may be impossible to achieve, it is still quite possible to reduce racial disparities in risk assessment. Indeed, AI-based approaches have shown great promise on this front.[11–13]

Hogan *et al.*[1] recommend that those developing risk assessments limit themselves to inputs that are "psychologically and theoretically" meaningful to avoid reflecting implicit and explicit racial biases. This assertion assumes that explicit and implicit racial biases are not reflected in psychologically and theoretically meaningful constructs and that careful human judgment minimizes racial biases. We argue that this view is mistaken. The fact that a variable is psychologically meaningful or theoretically important does not exclude the possibility that said variable is influenced by racial bias. Take the example of "attitudes that condone violence" (Ref. 1, p 8); Hogan and colleagues view this as a psychologically meaningful variable, and, therefore, racial disparities that result from this variable would be defensible. It seems that this variable is as susceptible as any other to racial biases, however, so we are unclear as to why racial biases that result from consideration of this sort of variable should be treated differently to any other.

Finally, we come to the question of transparency. Hogan *et al.* argue that, "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" (Ref. 1, p 6). This concern is quite common[2] and is often directed at machine-learning algorithms where the relationship between input and output is opaque. It ought to be noted that the factors underlying human judgments of risk are opaque as well. A clinician's judgment of risk can, and likely will, be shaped by processes that are either outside of conscious awareness or cannot be explained adequately. Of course, a clinician can formulate a "just-so" story to justify any risk-assessment classification, but the extent to which this explanation faithfully reflects the actual process by which the classification was reached is questionable. Unlike the human brain, AI-based risk assessments are based on mathematics, which permits us to ask and obtain answers to questions such as whether the risk classification would have been different if the person did not have a criminal record or was 10 years older. These sorts of answers are not reliable where the judgment has involved human discretion.

Over the coming years, we suspect that concerns about AI-based risk assessment will continue to be voiced. This is in no small part due to increasing enthusiasm about it, which in turn encourages unbridled speculation about unrealistic scenarios.[2] As work continues in this field, it is important that researchers maintain a balanced view about what exactly AI is, what it can do, and where the risks lie. Furthermore, as ethics concerns are raised, we ought to avoid the temptation to evaluate the drawbacks of AI without considering whether the available alternatives provide any improvements in these areas (or make things worse).

References

1. Hogan NR, Davidge EQ, Cor bian G: On the ethics and practicalities of artificial intelligence, risk assessment and race. J Am Acad Psychiatry Law 49(3) online, 2021. DOI:10.29158/JAAPL.200116-20
2. Berk R: Artificial intelligence, predictive policing and risk assessment for law enforcement. Ann Rev Criminology 4:209–37, 2021
3. Hayward KJ, Maas MM: Artificial intelligence and crime: a primer for criminologists. Crime Media Culture 17:209–233, 2021
4. Cockerill RG: Ethics implications of the use of artificial intelligence in violence risk assessment. J Am Acad Psychiatry Law 48:345–9, 2021
5. Spivak BL, Shepherd SM: Machine learning and forensic risk assessment: new frontiers. J Forensic Psychiatr Psychol 31:571–81, 2020
6. Bonta J: Risk-needs assessment and treatment. Choosing Correctional Options That Work: Defining the Demand and Evaluating the Supply. Edited by Harland AT. Thousand Oaks, CA: Sage, 1996, 18-32
7. Elyounes DA: Bail or jail? Judicial versus algorithmic decision-making in the pretrial system. Sci & Tech L Rev 21:376–445, 2020
8. Pepe M: The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford: Oxford University Press; 2003
9. Chouldechova A: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. J Big Data 5:153–63, 2017
10. Kleinberg J, Mullainathan S, Raghavan M: Inherent trade-offs in the fair determination of risk scores. arXiv, 2016. Available at: https://arxiv.org/pdf/1609.05807v1.pdf. Accessed March 17, 2021
11. Berk R, Heidari H, Jabbari S, *et al*: Fairness in criminal justice risk assessments: the state of the art. Sociological Methods Res 50:3–44, 2021
12. Coston A, Mishler A, Kennedy EH, Chouldechova A: Counterfactual risk assessments, evaluation and fairness. Proceedings of the 2020 Conference on Fairness, Accountability and Transparency. 2020
13. Mishler A, Kennedy EH, Chouldechova A: Fairness in risk assessment instruments: post-processing to achieve counterfactual equalized odds. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.