Ethics Implications of the Use of Artificial Intelligence in Violence Risk Assessment

Richard G. Cockerill, MD, MBE

Artificial intelligence is rapidly transforming the landscape of medicine. Specifically, algorithms powered by deep learning are already gaining increasingly wide adoption in fields such as radiology, pathology, and preventive medicine. Forensic psychiatry is a complex and intricate specialty that seeks to balance the disparate approaches of psychiatric science, which strives to explain human behavior deterministically, and the law, which emphasizes free choice and moral responsibility. This balancing, a central task of the forensic psychiatrist, is necessarily fraught with ambiguity. Such a complex task may intuitively seem impenetrable to artificial intelligence. This article first aims to challenge this assumption and then seeks to address the unique concerns posed by the adoption of artificial intelligence in violence risk assessment and prediction. The relevant ethics concerns are analyzed within the framework of traditional bioethics principles. Finally, recommendations for practitioners, ethicists, and others are offered as a starting point for further discussion.

J Am Acad Psychiatry Law 48:345–49, 2020. DOI:10.29158/JAAPL.003940-20

Reliable and accurate assessment of violence risk remains an elusive goal for forensic psychiatrists. Though numerous violence risk assessment tools exist, they have modest predictive value at best, and no single tool has gained universal adoption.[1] Thus, whether in the emergency room or in the court room, violence risk assessment remains largely the domain of expert opinion. Such opinions can appear subjective to lay people, may increase the risk that practitioner bias is introduced into an opinion, and are often difficult for judges, jurors, and lawyers to interpret and review.

These challenges present an opportunity for artificial intelligence (AI) and machine learning. AI is a broad category that encompasses any computer algorithm capable of performing some function previously thought to be exclusive to human intelligence. Machine learning describes the process by which these algorithms improve over time. These powerful tools are rapidly transforming the technological landscape of the world, leading some to believe we are on the verge of a fourth industrial revolution.[2] Critically, this industrial revolution is distinguished from previous ones by the type of tasks being automated. In the past, automation primarily affected physical tasks (e.g., auto assembly), but this new industrial revolution stands to transform how cognitive tasks are performed. AI-operated self-driving cars must not only be able to mechanically operate a motor vehicle but must also be able to perform high-level cognitive functions, such as predicting the behavior of a child playing in the road in its path to avoid a potentially catastrophic collision. Medicine, too, stands to be transformed by this technology.

Deep learning is a specific type of machine learning that utilizes artificial neural networks. This process often involves little human supervision. For example, an algorithm may be given thousands of images and told to identify lizards. It will only be told whether it is correct, but it will have received no instructions on what features define a lizard. In a 2018 study, a deep learning algorithm, CheXNeXt, was directly compared with practicing radiologists in

Published online May 14, 2020.

Dr. Cockerill is a Fellow in Forensic Psychiatry, UCLA-Semel Institute for Neuroscience and Behavior, Los Angeles, California. Address correspondence to: Richard G. Cockerill, MD, MBE, UCLA-Semel Institute for Neuroscience and Behavior, 760 Westwood Plaza, Los Angeles, CA 90095. E-mail: rcockerill@mednet.ucla.edu.

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

the interpretation of chest x-rays. The study indicated that the algorithm performed at a level comparable with that of practicing radiologists with an average of 12 years of experience. Perhaps more importantly, CheXNeXt was dramatically more efficient than its human competitors, requiring an average of 1.5 minutes to interpret 420 images compared with the 240 minutes required by the radiologists.[3]

Returning to forensic psychiatry, violence risk presents a problem that is notoriously difficult to assess, in no small part because of the myriad potentially relevant variables that must be considered in each individual case. Existing violence risk assessment tools, which apply a fixed set of risk factors to a population, are incapable of accounting for this complexity. Machine learning and, more specifically, deep learning bring something new to the task. When algorithms utilize machine learning, new data are constantly incorporated to improve and refine a predictive model. Stated simply, correct predictions reinforce the model, while incorrect predictions cause it to recalibrate. Therefore, whereas existing violence risk assessment tools are static, algorithms driven by deep learning are dynamic.

This advantage becomes even more pronounced when studying especially rare events, such as spree killings. A forensic psychiatrist may evaluate fewer than 10 spree killers in an entire career. An AI algorithm, on the other hand, could "observe" every spree killer active in the United States, perhaps seeing dozens of cases in any single year, and can refine its algorithm with every additional data point collected. Thus, a program utilizing machine learning may initially be equal or inferior to existing tools (or practitioners) but stands to make enormous strides over time.

Examining the CheXNeXt example again, it is not hard to imagine that this system may be capable of superior performance compared with practitioners in the very near future, which itself may provide enormous benefits, such as increased detection of occult cancers on chest x-rays, but also will present new ethics challenges. The dilemma presented when a practitioner "overrules" an algorithm (with generally superior performance) in evaluating a radiologic study deserves further consideration beyond the scope of this article.

It is clear that AI already has an important role to play in medical diagnostics, one that is only likely to increase as the technology continues to improve. Assessing future violence risk in a person convicted of a felony, however, is a much different challenge than reading a chest x-ray. Is the technology up to this daunting task, or is it years away from feasible implementation? The future may be closer than we think. According to a February 2019 BBC report, 14 U.K. police forces had already started to use "crime-prediction software" in their crime-prevention efforts.[4] Critically, these technologies rely heavily on machine learning. Thus far, there are two types of software available for law enforcement. One type, predictive mapping, assesses the likelihood of certain crimes occurring in specific locations. Police can then proactively increase their presence in those places to prevent such crimes. Predictive mapping has been applied to terrorism,[5] school safety,[6] gang activity,[7] and gun violence.[8] The other type of software generates individual risk scores that assess a person's risk of engaging in future criminal activity. Typically this involves those on parole or probation. Those identified as high risk would be targeted with preventive interventions or increased surveillance. One such program, known as Operation LASER, had been implemented in Los Angeles by the Los Angeles Police Department, but it was discontinued in early 2019 in response to concerns about privacy and racial bias.[9] Such applications of AI in policing, and their unintended consequences, are becoming eerily similar to what was recently the domain of science fiction. The widely acclaimed J.J. Abrams series, Person of Interest,[10] which features a terrorism-fighting supercomputer called "The Machine" appears particularly prescient now.

AI is also coming to violence risk assessment on inpatient wards. A recent study in the Netherlands applied machine learning techniques to a large data set of clinical notes to predict future violent behavior among psychiatric inpatients; the results showed generally equal or superior predictive value when compared with existing tools. Like CheXNeXt, this algorithm relied on deep learning and thus could formulate predictions based on its unsupervised interpretation of clinical notes and will continue to improve as it is exposed to more data.[11] Finally, in a closely related area, Facebook has quietly rolled out its machine learning–driven suicide prevention tool, which analyzes an enormous wealth of user data to assess suicide risk. As of December 2018, this had resulted in at least 3,500 notifications to local emergency services in the United States alone.[12] Though

the company has not published outcome data for this program, it strains credulity to imagine there have not been some false positive reports. The potential consequences of such mistakes, including significant violations of privacy and liberty, are difficult to understate.

AI in general, and deep learning specifically, clearly have enormous potential in their application to forensic psychiatry and violence risk assessments. These technologies also come with significant risks. The stakes are incredibly high; "smart" algorithms stand to play a critical role in decisions to involuntarily commit or medicate individuals, the provision of sentencing recommendations, and even the guidance of targeted police surveillance. The values systems that guide these algorithms, for better or worse, will be determined by those that design them. It is therefore critical to identify the ethics implications posed by the use of this technology, both to guide its designers and those who interpret its work.

Analysis

The ethics concerns of interest can be well examined using the approach developed by Beauchamp and Childress, emphasizing autonomy, beneficence, nonmaleficence, and justice as guiding principles.[13]

Autonomy

The adoption of AI-driven tools in violence prediction and prevention has obvious implications for personal freedom and autonomy. Looking at crime-prediction software and Facebook's suicide-prevention efforts as just two examples, it becomes evident that the data gathered, which may include web search histories, social media posts, online shopping records, and, in the future, even personal biometric data (e.g., from wearable devices), is generally thought of as at least somewhat private by consumers. Personal health information, protected by HIPAA, presents particular concerns. HIPAA contains a "serious and imminent threat" exception to privacy rules that allows patient records to be violated to prevent some serious harm. Could an algorithm's initial assessment of an individual's violence risk, based on publicly available data, be used as a pretext to access that individual's health records? Freedom of thought, speech, and expression are also directly implicated by this technology. Detaining someone before that person has committed a violent act is not unprecedented (i.e., psychiatric holds), but it cannot be taken lightly.

To illustrate this point, consider the hypothetical case of Kyle, a 19-year-old man with an avid interest in horror films. Kyle has no criminal history but does have a history of depression, for which he takes fluoxetine from his primary care doctor. After watching "The Silence of the Lambs" for the first time, Kyle develops a morbid fascination with cannibalism and begins avidly searching online for more information about it, including visiting websites describing how to complete such acts and "get away with it," as well as fan sites for Jeffrey Dahmer. At the same time, his smart watch detects elevated pulse rate, blood pressure, and respiratory rate. This, taken with ample additional data about Kyle's recent activities, social isolation, and psychiatric risk factors, is enough for a hypothetical "violence prevention algorithm" to flag him as high risk for imminent violence, and emergency services are contacted. As a result, he is eventually detained and psychiatrically hospitalized for 30 days, all the while denying any intent to commit any acts of violence. After discharge, Kyle continues to deny any violent intent during this period and decries this episode as a terrible injustice and violation of his liberty.

Beneficence and Nonmaleficence

It is reasonable to consider these principles together because, especially in this case, they are inextricably linked. That is, the benefits of using AI algorithms to prevent violence are directly related to its costs. The more we sacrifice privacy, the more data are available to the algorithms. More data means a better, smarter algorithm. A better algorithm means better results and potentially fewer violent tragedies. The inverse of this is true as well. Keeping more of our data private preserves liberty, but it also has a cost. We then have weaker tools with which to prevent violence and are left with more preventable tragedies. Balancing these equities in the right way is of enormous social importance. To further elucidate these concerns, let us return to the hypothetical case of Kyle. Imagine the scenario described above developed quite differently. Say, after observing Kyle's online behavior and biometric data as above, the algorithm identifies him as having a 97 percent risk for an "imminent violent act" within the next week. A human reviewer receives this report and reviews the data. Being familiar "The Silence of the Lambs," the

reviewer identifies the pattern of behavior as that of another horror enthusiast and overrules the algorithm. Two days later, Sheila, a 12-year-old girl and neighbor of Kyle, is murdered by him in a seemingly random act of violence. Sheila's family becomes aware that Kyle was identified as high risk by the algorithm but that this was not reported to the authorities. Does her family have a legitimate legal or ethics claim against the company? It would certainly seem that such an argument could be made. Thus, when such AI algorithms have sufficiently good predictive power, those who can use them may have a duty to do so, or at least to provide their analyses to relevant parties. This would not be dissimilar to psychiatry's *Tarasoff* rule, to mandated reports of suspected child abuse by health care professionals, or to legal commitment procedures for those deemed at risk to harm others, except that in the case of AI-driven algorithms, the assessments may be much more reliable. How far such a duty is extended will have great consequences. Sex offenders, for example, are closely surveilled when released from prison. In this case, society has placed great weight on the prevention of the sexual abuse of children, to the extent that it is willing to tolerate great intrusions on the privacy of these offenders, who may be required to submit to electronic surveillance of essentially all of their online activity. It is conceivable that similar arrangements may be proposed for those deemed at especially high risk for the commitment of violent crime. In these cases, too, the harms of such surveillance must be carefully balanced with its benefits.

Justice

For the purposes of this discussion, justice refers to fairness and equal treatment under the law. AI-driven algorithms could have enormous positive impact when applied to the inequities present in the justice system. It is well known that black and Latino men face disproportionately long sentences for similar crimes in the United States.[14] This bias almost certainly pertains to clinician violence risk assessments. A model that is truly neutral could play a vital role in moving toward a more just legal system.

There are risks, however, that may not be readily apparent. Smart algorithms, although powerful tools, are still only as good as the data they analyze. For example, some existing sentencing recommendation algorithms used by courts around the world, while excluding race and ethnicity as input data, still appear to be biased toward harsher penalties for black men, although this has been debated.[15,16] Dealing with matters of justice and fairness, especially as they relate to race, is especially challenging for those designing AI systems. It is difficult to argue that racial justice is a problem that our society fully understands, much less knows how to solve, yet this is precisely the territory that algorithms are increasingly entering. These tools could be enormously effective in the development of a more equitable system of justice, but we must first decide how such a system should look.

Discussion and Recommendations

There is little question that AI will have a major impact on the interface between psychiatry and law. This article can only hope to introduce the outlines of the technology and its ethics implications. The potential for rapid adoption of such powerful tools shines a bright spotlight on the ethics concerns, which must now be analyzed with heightened urgency. AI algorithms are already making ethics judgments; as they become more powerful and widely adopted, it is critical that such judgment be appropriate.

In the coming years, forensic psychiatrists will be increasingly asked to interpret data from smart algorithms in making recommendations involving competency, commitment, sentencing decisions, the use of involuntary medications, and more. To do this, they must have some level of understanding of the technology and the ethics questions implicated by it. To this end, the following recommendations may be considered a starting point for addressing this challenge:

> Basic instruction in computer science and its interaction with medicine and psychiatry should be offered in medical school, residency, and forensic psychiatry fellowship training programs.
>
> Forensic psychiatry programs, in particular, should offer training in the admissibility and appropriate use of digital evidence, including biodata, in legal settings.
>
> Forensic psychiatrists, armed with this knowledge, can then aim to be translators of information provided by algorithms to attorneys, judges, and jurors, also paying heed to the ethics concerns raised by its use.

Interdisciplinary programs between computer scientists, ethicists, and forensic psychiatrists should be developed and strengthened to increase understanding of the potential value and risks posed by this technology.

Finally, forensic psychiatrists should play a key role in informing policy regarding the use of such technology, specifically addressing questions about the detainment and appropriate treatment of at-risk individuals identified by such technology.

Acknowledgments

Special thanks to Dr. Robert Weinstock, UCLA Forensic Psychiatry Program Director, for his invaluable feedback and guidance in the development of this manuscript.

References

1. Douglas T, Pugh J, Singh I, *et al*: Risk assessment tools in criminal justice and forensic psychiatry: the need for better data. Eur Psychiatry 42:134–7, 2017
2. Xu M, David JM, Kim SH: The fourth industrial revolution: opportunities and challenges. Int J Fin Res 9:90–5, 2018
3. 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
4. Kelion L: Crime prediction software 'adopted by 14 UK police forces.' BBC News. February 4, 2019. Available at: https://www.bbc.com/news/technology-47118229. Accessed December 31, 2019
5. Ding F, Ge Q, Jiang D, *et al*: Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approaches. PLoS One 12:e0179057, 2017
6. Barzman D, Ni Y, Griffey M, *et al*: Automated risk assessment for school violence: a piloted study. Psychiatr Q 89:817–28, 2018
7. McCullom R: A murdered teen, two million tweets and an experiment to fight gun violence. Nature News. September 4, 2018. Available at: https://www.nature.com/articles/d41586-018-06169-8. Accessed January 2, 2020
8. Goin DE, Rudolph KE, Ahern J: Predictors of firearm violence in urban communities: a machine-learning approach. Health Place 51:61–7, 2018
9. Puente M: LAPD ends another data-driven crime program touted to target violent offenders. Los Angeles Times. April 12, 2019. Available at: https://www.latimes.com/local/lanow/la-me-laser-lapd-crime-data-program-20190412-story.html. Accessed January 2, 2020
10. Person of Interest. Produced by Nolan J, Plageman G, Abrams JJ, *et al*. Los Angeles: Warner Bros. Television, 2011–2016
11. Menger V, Spruit M, van Est R, *et al*: Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. JAMA Netw Open 2:e196709, 2019
12. Singer N: In screening for suicide risk, Facebook takes on tricky public health role. New York Times. December 31, 2018:A1
13. Beauchamp TL, Childress JF: Principles of Biomedical Ethics, Seventh Edition. New York: Oxford University Press, 2013
14. Sutton JR: Structural bias in the sentencing of felony defendants. Soc Sci Res 42:1207–21, 2013
15. van Eijk G: Socioeconomic marginality in sentencing: the built-in bias in risk assessment tools and the reproduction of social inequality. Punish Soc 19:463–81, 2017
16. Washington AL: How to argue with an algorithm: lessons from the COMPAS ProPublica Debate. Colo Tech L J 17:131–60, 2018