刑事司法科学与治理
*Criminal Justice Science & Governance*
2023 年 第 4 卷 第 2 期

# 法医精神病学和刑事司法中的神经预测和人工智能：神经法视角

Leda Tortora[1] Gerben Meynen[2,3] Johannes Bijlsma[2]
Enrico Tronci[4] Stefano Ferracuti[1] 著 万 欢[1] 译

1．罗马第一大学人类神经科学系，罗马；
2．乌得勒支大学 Willem Pompe 刑法和犯罪学研究所 /Utrecht 问责和责任法中心（UCALL），乌得勒支；
3．阿姆斯特丹自由大学人文学院，阿姆斯特丹；
4．罗马第一大学计算机科学系，罗马

**摘 要** | 神经影像学与人工智能结合使用的进步，特别是机器学习技术的使用，促进了大脑阅读技术的发展。这些技术在不久的将来可能被广泛应用，例如测谎、神经营销或大脑计算机接口。其中一些原则上也可用于法医精神病学。这些方法在法医精神病学中的应用可能有助于提高风险评估的准确性并确定可能的干预措施。这种技术可以称为“人工智能神经预测”，它涉及识别潜在的神经认知标志物来预测累犯。然而，这项技术的未来意义以及神经科学和人工智能在暴力风险评估中的作用仍有待确定。本文回顾和分析了有关使用大脑阅读人工智能对暴力和再逮捕进行神经预测的文献，以确定未来在法医精神病学和刑事司法领域使用这些技术的可能性和挑战，同时考虑了法律影响和伦理问题。分析表明，需要对人工智能神经预测技术进行更多研究，并且非常有必要了解如何在法医精神病学领域的风险评估中实施这些技术。除了人工智能神经预测的诱人潜力之外，本文认为，在这些技术完全可用时以及在其研究和开发过程中都应该对其在刑事司法和法医精神病学中的使用进行彻底的危害或者益处分析。

**关键词** | 神经预测；人工智能；累犯；法医精神病学；风险评估；神经法



## 一、引言

风险评估是刑事司法系统的重要组成部分。近年来，人们对开发新的工具和技术以改善法医精神病学和刑事司法领域的风险评估越来越感兴趣[1]。目前，已经开发了 200 多种暴力风险评估工具，通

［1］万欢，中南财经政法大学刑事司法学院2021级博士研究生。
［2］Monahan J，Skeem J L. Risk Assessment in Criminal Sentencing［J］. Annual Review of Clinical Psychology，2015：53.

常是集成的临床精算工具，用于预测暴力、反社会性和性行为[1]，并且它们在刑事司法环境中的使用似乎显著增加。这些方法的中心目标是正确识别高风险和低风险罪犯。根据司法管辖权的不同，它们被用于为一系列医疗法律决定提供信息，例如量刑、假释、民事承诺、死刑、少年法庭的处置以及发现精神错乱后的释放问题[2]。近年来，人工智能（AI）被用于提高风险评估的预测准确性。

算法风险评估的使用随着神经影像学领域的研究而发展，推动了“大脑阅读”技术的发展，这些技术在一定程度上能够根据人的大脑活动解码心理状态[3]或者根据人的大脑结构和功能将他们分组[4]。该技术的一个可能的司法应用是识别危险的罪犯。人工智能和神经影像学的结合促进了所谓的“人工智能神经预测”的发展，即使用结构或功能性大脑参数与机器学习方法相结合来进行临床或行为预测。也许在不久的将来，人工智能神经预测可以更普遍地用于预测法医精神病学和刑事司法中的累犯风险。然而，这类技术的应用引发了法律和伦理问题。

本文的目的是确定未来在法医精神病学和刑事司法领域使用人工智能神经预测暴力和累犯的可能性和挑战，并讨论法律影响和伦理问题。在第二部分，我们将讨论风险评估技术。在第三部分，我们探讨了当前使用神经影像与人工智能相结合的“大脑阅读”技术。在第四部分，我们概述了近年使用神经影像数据与人工智能相结合的神经预测来预测累犯的研究。在第五部分，我们讨论了预测分析的技术限制和陷阱。最后，在第六部分，我们讨论了应用这些技术所引发的伦理和法律问题。

## 二、风险评估：最先进的技术

在过去的二十年中，美国和欧洲对暴力风险评估工具的兴趣和研究显著增加，提出了不同的方法，从基于回归的严格精算工具到算法风险评估，提供重新犯罪的概率估计以及结构化的专业判断[5][6]。最初，精算方法在该领域占主导地位，它们虽然有一定的预测价值，但是仍然非常有限[7]。

与个人暴力或攻击性行为的可能性增加相关的风险变量包括犯罪需求（增加累犯风险的个人特征）、人口统计、社会经济地位和智力[8]。风险因素通常分为静态因素和动态因素。静态因素是历史性的，不会改变（如犯罪历史、犯罪类型、童年虐待）；动态因素原则上是可变的，因此它们提供了干预的机会，可以改变未来的风险（如冲动、吸毒、社会支持、工作、治疗依从性）。一些动态因素相当稳定，而另一些则更“流动”。动态因素需要多次测量，有时测量的间隔期很短。

然而，目前风险评估工具的结果远未达到完美的程度，尤其是对于长期预测而言。当前的刑事风

[1] Singh J P, Desmarais S L, Hurducas C, et al. International perspectives on the practical application of violence risk assessment: a global survey of 44 countries [J]. International Journal of Forensic Mental Health, 2014, 13(3): 193-206.

[2] Conroy M A, Murrie D C. Forensic Assessment of Violence Risk: A Guide for Risk Assessment and Risk Management [M]. Hoboken: John Wiley & Sons Inc, 2007.

[3] Haynes J D, Rees G. Neuroimaging: decoding mental states from brain activity in humans [J]. Nature reviews neuroscience, 2006(7): 523-534.

[4] Koutsouleris N, Borgwardt S, Meisenzahl E M, et al. Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: results from the FePsy study [J]. Schizophrenia Bulletin, 2012, 38(6): 1234-1246.

[5] Hart S D. The role of psychopathy in assessing risk for violence: conceptual and methodological issues [J]. Legal and Criminological Psychology, 1998, 3: 121-137.

[6] Douglas K S, Kropp P K. A prevention-based paradigm for violence risk assessment: clinical and research applications [J]. Criminal Justice & Behavior, 2002, 29(5): 617-658.

[7] Fazel S, Singh J P, Doll H, et al. Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24827 people: systematic review and meta-analysis [J]. British Medical Association, 2012, 345: e4692.

[8] Gendreau P, Little T, Goggin C. A meta-analysis of the predictors of adult offender recidivism: what works [J]. Criminology, 1996, 34(4): 575-608.

险评估工具显示出较差到中等的准确性，需要考虑在假阳性和假阴性之间取得良好的平衡，这取决于社会和政治背景以及使用该工具时所处的刑事司法程序的阶段[1]。通常，当风险评估工具将个人归类为低风险时，通常是正确的。然而，如果该工具将某人归类为高风险，这通常是不正确的，几乎超过一半的高风险人群被错误地归类。假阳性（预计被告会再次犯罪，但事实并非如此）似乎比假阴性（预计被告不会再次犯罪，但确实如此）更常见[2]。

这往往导致许多人可能被监禁或继续被监禁，而他们实际上不会对社会构成危险。Fazel 等人（2012）指出："这些发现的一个含义是，即使经过 30 年的发展，在大多数情况下可以预测暴力、性或犯罪风险的观点也不是基于证据的。"这种对当前事态的诊断使得寻找改进法医精神病学和刑事司法风险评估的方法变得非常重要。

与通常来源于各种形式的回归分析的经典方法相比，算法有望对犯罪行为进行更准确的预测[3]。它们可用于为未来的暴力提供个体化风险措施，并有助于制定预防和治疗决策，以尽量减少风险因素并突出保护因素。包含机器学习的风险评估工具已用于审前风险评估、量刑和康复[4]，并且可能在司法决策中发挥重要作用，以指导"关于保释、缓刑 / 假释的决定、法院命令的治疗和民事承诺"[5]。

## 三、人工智能和神经影像学

脑成像技术的快速发展以及人工智能技术在社会诸多领域中的影响力日益增强，从社交网络到医疗保健和警察政策[6]，引起了人们对脑成像与人工智能相结合以改善对未来暴力行为的风险评估和预测的潜在用途的兴趣。

在过去十年中，非侵入性解剖和功能性神经成像技术取得了重大发展，产生了大量数据。统计机器学习方法有助于以越来越高的精度分析大量神经数据[7]和高维数据集建模[8]。将统计机器学习方法应用于神经影像数据被称为多体素模式分析（MVPA）[9]。与一次仅分析一个位置的传统单变量方法不同，这些方法允许识别数据中的空间和时间模式，区分认知任务或具有更高灵敏度的主题组，共同分析来自区域内单个体素的数据[10]。

自 MVPA 方法问世以来，该方法已成为"健康和临床人群的神经影像学"中的一种流行方法。


[1] Douglas T，Pugh J，Singh I，et al. Risk assessment tools in criminal justice and forensic psychiatry：the need for better data [J]. European Psychiatry，2017，42：134-137.

[2] Fazel S，Singh J P，Doll H，et al. Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24827 people：systematic review and meta-analysis [J]. British Medical Association，2012，345：e4692.

[3] Berk R，Hyatt J. Machine learning forecasts of risk to inform sentencing decisions [J]. Federal Sentencing Reporter，2015，27：222-228.

[4] Kehl D，Guo P，Kessler S. Algorithms in the Criminal Justice System：Assessing the Use of Risk Assessments in Sentencing [J]. Dash. Harvard. Edu，2017.

[5] Poldrack R A，Monahan J，Imrey P B，et al. Predicting violent behavior：what can neuroscience add? [J]. Trends in Cognitive Sciences，2018，22：111-123.

[6] Berk R，Heidari H，Jabbari S，et al. Fairness in criminal justice risk assessments：the state of the art [J] SAGE Publications，2018.

[7] Lemm S，Benjamin B，Thorsten D，et al. Introduction to machine learning for brain imaging [J]. Neuroimage，2011，56：387-399.

[8] Abraham A，Pedregosa F，Eickenberg M，et al. Machine learning for neuroimaging with scikit-learn [J]. Frontiers in Neuroinformatics，2014，8：14.

[9] Ombao H，Lindquist M，Thompson W，et al. Handbook of Neuroimaging Data Analysis [M]. New York：Chapman and Hall/CRC，2017：164-169.

[10] Haynes J D，Rees G. Neuroimaging：decoding mental states from brain activity in humans [J]. Nature reviews neuroscience，2006（7）：523-534.

研究表明，神经影像数据中存在的信息可用于解码，在一定程度上意图和感知状态，以及区分健康和患病的大脑[1]。MVPA已被应用于解码视觉特征，如边缘方向[2]、执行一项任务而不是另一项任务的意图[3]、任务准备的顺序阶段[4]，和测谎[5][6]。虽然传统的功能成像研究比较了不同实验条件下的大脑活动，以确定哪些大脑区域被特定任务激活，但MVPA在大脑阅读中的应用使用“大脑活动模式来执行反向推理并决定受试者在看什么或在想什么”[1][7]。

这些技术可以被认为是“读脑”或“读心”技术，它们将统计机器学习方法与神经影像数据相结合，以揭示有关大脑/心灵的信息。大脑阅读经常在视觉感知领域进行研究，其目的是展示大脑中的体验如何编码。研究人员最近成功训练了一个深度神经网络[8]执行来自大脑的视觉图像重建[9]，解码梦境的视觉内容[10]，并通过使用人工智能分析来自观看视频的受试者的fMRI扫描来解码大脑“看到”的内容[11]。尽管取得很多重要发现，但这些方法仍然显示出许多局限性，这使得“通用读心术”不太可能在不久的将来出现。尽管如此，简单的应用已经开始出现，包括脑机接口、测谎研究和神经营销领域的消费者决策预测方法[12]。

除了对精神状态的发生和性质进行推断[12]之外，MVPA技术的另一个应用领域是分类。例如，研究发现，MVPA技术可以通过基于大脑活动区分群体中的个体或基于识别大脑活动或结构模式的大脑数据将个体分类来预测疾病发作[13]。通过提取活动或结构异常的模式，可以将治疗反应者与无反应者区分开，这些模式或结构异常可预测异常认知发展，与神经影像数据的临床结果预测相关[1]。一些模型用于区分临床群体，如阿尔茨海默病患者

[1] Bray S，Chang C，Hoeft F. Applications of multivariate pattern classifification analyses in developmental neuroimaging of healthy and clinical populations [J]. Frontiers in Human Neuroscience，2009，3：32.

[2] Kamitani Y，Tong F. Decoding the visual and subjective contents of the human brain [J]. Nature Neuroscience，2005，8：679-685.

[3] Haynes J D，Sakai K，Rees G，et al. Reading hidden intentions in the human brain [J]. Current Biology，2007，17（4）：323-328.

[4] Bode S，Haynes J D. Decoding sequential stages of task preparation in the human brain [J]. Neuroimage，2009，45：606-613.

[5] Davatzikos C，Ruparel K，Fan Y，et al. Classifying spatial patterns of brain activity with machine learning methods：application to lie detection [J]. Neuroimage，2005，28（3）：663-668.

[6] Blitz M J. Lie detection，mind reading，and brain reading [M] //Palgrave Studies in Law，Neuroscience，and Human Behavior. Cham：Palgrave Macmillan，2017：45-58.

[7] Cox D D，Savoy R L. Functional magnetic resonance imaging（fMRI）brain reading：detecting and classifying distributed patterns of fMRI activity in human visual cortex [J]. Neuroimage，2003，19（2）：261-270.

[8] 神经网络是“由许多并行运行的简单处理元素组成的系统，其功能取决于网络结构、连接强度以及在计算元素或节点处执行的处理。”

[9] Shen G，Horikawa T，Majima K，et al. Deep image reconstruction from human brain activity [J]. PLoS Computational Biology，2019，15：e1006633.

[10] Horikawa T，Tamaki M，Miyawaki Y，et al. Neural decoding of visual imagery during sleep [J]. Science，2013，340：639-642.

[11] Wen H，Shi J，Zhang Y，et al. Neural encoding and decoding with deep learning for dynamic natural vision [J]. Cerebral Cortex，2017，28：4136-4160.

[12] Haynes J D. Brain reading [M] //I Know What You're Thinking：Brain imaging and Mental Privacy. Oxford：Oxford University Press，2012：29-40.

[13] Koutsouleris N，Borgwardt S，Meisenzahl E M，et al. Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers：results from the FePsy study [J]. Schizophrenia Bulletin，2012，38（6）：1234-1246.

和认知正常的老年人[1]、帕金森病患者和健康对照者[2]、精神分裂症患者和健康对照者[3]，或检测大脑功能障碍，如自闭症和注意力缺陷多动障碍（ADHD）[4][5]，并区分人格特征的水平，如精神病[6]。

关于成瘾结果的预测也有相关的研究。机器学习分类器能够使用事件相关电位（ERP）[7][8]和分析 fMRI 数据的功能网络连接（FNC）来预测监狱囚犯群体的药物滥用治疗完成情况[9]。此外，使用最近开发的机器学习方法 CPM[10]可以识别“神经指纹”来预测治疗期间的可卡因戒断情况。

## 四、累犯的人工智能神经预测

行为特征与人类大脑的特征相关，有时甚至是显著相关，这为开发预测算法提供了新的可能性，有助于预测个体的性格。这些方法被称为“神经预测”，即使用结构或功能性大脑变量来预测预后、治疗结果和行为预测[11]。尽管目前听起来像是科幻小说，但随着非侵入性神经成像技术的不断发展以及算法计算能力的增长，人工智能对累犯的神经预测很可能在不久的将来成为现实。

虽然仍然需要收集“犯罪”大脑的生物标志物，但神经犯罪学领域的研究普遍集中在分析主要特征为持续反社会行为的人格障碍的结构和功能神经标志物，例如反社会性人格障碍（ASPD）[12]和精神病[13]，因为它们似乎与高累犯率相关性最高。研究表明，这些特定的临床人群具有许多共同特征，例如行为去抑制或缺乏同理心，这些特征应该具有共同的神经生物学基础[14]。

---


[1] Klöppel S，Stonnington C M，Chu C，et al. Automatic classifification of MR scans in Alzheimer's disease [J]. Brain，2013，131（3）：681-689.

[2] Rubbert C，Mathys C，Jockwitz C，et al. Machine-learning identififfies Parkinson's disease patients based on resting-state between-network functional connectivity [J]. British Journal of Radiology. 2019：20180886.

[3] Kim J，Calhoun V D，Shim E，et al. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classifification performance：evidence from whole-brain restingstate functional connectivity patterns of schizophrenia [J]. Neuroimage，2016，124（A）：127-146.

[4] Heinsfeld A S，Franco A R，Craddock R C，et al. Identifification of autism spectrum disorder using deep learning and the ABIDE dataset [J]. Neuroimage Clinical，2018，17：16-23.

[5] Sen B，Borle N C，Greiner R，et al. A general prediction model for the detection of ADHD and Autism using structural and functional M. R. I [J]. PloS one，2018，13：e0194856.

[6] Steele V R，Rao V，Calhoun V D，et al. Machine learning of structural magnetic resonance imaging predicts psychopathic traits in adolescent offffenders [J]. Neuroimage，2015：265-273.

[7] Steele V R，Fink B C，Maurer J M，et al. Brain potentials measured during a Go/NoGo task predict completion of substance abuse treatment [J]. Biological Psychiatry，2014，76（1）：75-83.

[8] Fink B C，Steele V R，Maurer M J，et al. Brain potentials predict substance abuse treatment completion in a prison sample [J]. Brain and Behavior，2016，6：501.

[9] Steele V R，Maurer J M，Arbabshirani M R，et al. Machine learning of functional magnetic resonance imaging network connectivity predicts substance abuse treatment completion [J]. Biological Psychiatry，2018，3：141-149.

[10] Yip S W，Scheinost D，Potenza M N，et al. Connectomebased prediction of cocaine abstinence [J]. American journal of psychiatry，2019，176：156-164.

[11] Morse S J. Neuroprediction：New Technology，Old Problems [J]. Faculty Scholarship at Penn Law，2015：1619.

[12] De Brito S A，Mechelli A，Wilke M，et al. Size matters：increased grey matter in boys with conduct problems and callous-unemotional traits [J]. Brain，2009，132（4）：843-852.

[13] Umbach R，Berryessa C，Raine A. Brain imaging research on psychopathy：implications for punishment，prediction，and treatment in youth and adults [J]. Journal of Criminal Justice，2015，43：295-306.

[14] Coppola F. Mapping the brain to predict antisocial behaviour：new frontiers in neurocriminology，“new” challenges for criminal justice [J]. UCL Journal of Law and Jurisprudence-Special Issue，2018，1：106-110.

例如，在具有精神病特征的个体中观察到边缘和旁边缘区域的异常[1]；与前额叶皮层相关的损伤与去抑制、情绪不稳定和冲动有关[2][3]。

尽管如此，从使用传统方法获得的所有关于神经犯罪学的相关研究成果中可以发现，目前尚不能预测未来的风险。然而，在人工智能预测模型中加入神经数据似乎提供了可能性。

Aharoni 等人（2013）进行的一项研究迈出了使用神经影像数据进行人工智能预测模型的第一步，他们使用 fMRI 数据来预测累犯。研究发现，在执行 / 不执行任务期间，背侧前扣带皮层（dACC）是一个与冲动控制和错误处理相关的大脑区域，其激活似乎与重新停止有关。在保持所有其他风险因素不变的情况下，前扣带回活动相对较低的罪犯再次被捕的概率大约是该区域活动较多的罪犯的两倍。因此，低前扣带回活动可能是持续犯罪行为的潜在神经认知生物标志物[4]。

Kiehl 等人（2018）的一项研究将机器学习与神经影像学相结合来测试大脑年龄是否有助于预测再逮捕。实际年龄年轻被认为是累犯的关键风险因素之一。年轻的被告更有可能从事危险行为。他们还提出，与实际年龄相比，大脑年龄是解释个体差异的更好的衡量标准。研究结果表明，涉及大脑年龄神经测量的预测模型比以前仅包括心理和行为测量的模型表现得更好。

Delfifin 等人（2019）的一项研究表明，通过将神经影像数据纳入人工智能风险评估模型，可以改善法医精神病学的累犯预测。作者指出，在扩展的人工智能预测模型中包含静息状态区域脑血流量（rCBF）测量，该模型包含来自八个大脑区域的神经测量，在法医精神病患者的长期随访中，与传统的经验风险因素相比，预测性能有所提高。他们将“经典”风险评估与神经影像学相结合，发现在法医精神病人群中应用这种方式比单独使用经典因素能够进行更好的预测[5]。

综上所述，人工智能神经预测研究的初步发现已经产生了一些有希望的结果。尽管如此，在法医人群中使用人工智能和“大脑阅读”的可能性引起了一些道德和法律问题，刑事司法领域应该对它们的未来使用保持谨慎态度。

在维护罪犯个人权利和加强公共安全之间取得平衡至关重要。

## 五、预测分析：技术限制和陷阱

尽管前文已讨论了有关未来可能使用人工智能神经预测技术的机会，但仍应考虑一些限制。事实上，关于预测工具及其成功应用的研究仍然是一项具有挑战性的任务[6]。

将机器学习方法和基于神经影像学的脑疾病单学科预测相结合对患有异质性疾病的患者进行分类[7][8]的研究在计算精神病学领域众所周知。

[1] Anderson N E，Kiehl K A. The psychopath magnetized：insights from brain imaging [J]. Trends in Cognitive Sciences，2012，16：52–60.

[2] Chow T W. Personality in frontal lobe disorders [J]. Current Psychiatry Reports，2000，2：446–451.

[3] Yang Y，Raine A. Prefrontal structural and functional brain imaging fifindings in antisocial，violent，and psychopathic individuals：a meta–analysis [J]. Psychiatry Research，2009，174：81–88.

[4] Aharoni E，Vincent G M，Harenski C L，et al. Neuroprediction of future rearrest [J]. Proceedings of the national academy of sciences of the united states of america，2013.

[5] Delfifin C，Krona H，Andine P，et al. Prediction of recidivism in a long–term follow–up of forensic psychiatric patients：incremental effffects of neuroimaging data [J]. PLoS One，2019，14：e0217127.

[6] Poldrack R A，Huckins G，Varoquaux G. Establishment of best practices for evidence for prediction：a review [J]. JAMA Psychiatry，2019，27：2019.

[7] Arbabshirani M R，Plis S，Sui J，et al. Single subject prediction of brain disorders in neuroimaging：promises and pitfalls [J]. Neuroimage，2017，145（B）：137–165.

[8] Bzdok D，Meyer–Lindenberg A. Machine learning for precision psychiatry：opportunities and challenges [J]. Biological Psychiatry，2018，3：223–230.

这些研究报告了不同程度的准确性[1]，引发了人们对该方法的担忧[2]。事实上，预测建模需要最佳实践[3]；神经预测模型存在一个问题：即使它们可以管理复杂的数据，例如脑成像扫描，也需要最佳实践来确保具备足够的统计能力来测试其有效性[4]。以下是值得关注的问题。

首先，神经预测技术的应用需要从组级到个人预测的推断[5]。另一个挑战涉及在新组中验证结果——与用于训练算法的数据集不同。预测模型的有效性通过它们的泛化能力来评估。对于大多数学习算法，标准做法是通过称为“交叉验证”的过程来估计泛化性能：数据集分为两组，用于拟合模型的训练集和测试集[6]，并且数据的子集用于迭代地训练和测试模型的预测性能。

值得注意的是，对小样本使用交叉验证会导致预测准确性的高度可变和夸大估计[7]。训练机器学习算法需要大量数据，使用有限的样本量可能会导致所谓的过度拟合，其中模型完全适合用于训练它的特定数据集，但不适用于新的和未使用过的数据。关于数据集的合适的样本数量仍未达成一致，Luedtke 等人（2019）建议对不小于数百个观测值的样本进行预测分析。然而，获取大量样本通常既困难又昂贵，特别是在涉及神经影像数据时。

## 六、道德和法律挑战

使用人工智能神经预测技术预测累犯会引发伦理和法律问题，但也引发了新的可能性。在下文中，我们将讨论一些核心的伦理和法律问题。

首先，面临着偏见的问题。自算法风险评估出现以来，许多报告都记录了它们存在“危险”偏见的事实。ProPublica 于 2016 年 5 月报道了最著名的所谓人工智能偏见案例。根据 ProPublica 的说法，COMPAS 是一种在美国广泛使用的算法，通过预测重新犯罪的可能性来指导量刑，结果证明对黑人被告存在种族偏见，因为他们比白人被告更有可能被错误地归类为高风险（“误报”）[8][9]。最近，COMPAS 也被描述为“性别歧视算法”，因为它的算法结果似乎系统地将女性过度分类为高风险群体[10]。同样地，Predpol 是一种旨在预测犯罪发生时间和地点的算法，在对人权数据分析小组进行分析后，已于 2016 年在美国多个州使用，该算法被发现导致警察不公平地针对某些社区。警察被反复派

[1] Neuhaus A H，Popescu F C. Sample size，model robustness，and classification accuracy in diagnostic multivariate neuroimaging analyses [J]. Biological Psychiatry，2018，84：e0081-2.

[2] Cearns M，Hahn T，Baune B T. Recommendations and future directions for supervised machine learning in psychiatry [J]. Translational Psychiatry，2019，9：271.

[3] Poldrack R A，Huckins G，Varoquaux G. Establishment of best practices for evidence for prediction：a review [J]. JAMA Psychiatry，2019，27：2019.

[4] Varoquaux G. Cross-validation failure：small sample sizes lead to large error bars [J]. Neuroimage，2018，180（A）：68.

[5] Hahn T，Nierenberg A，Whitfifield-Gabrieli S. Predictive analytics in mental health：applications，guidelines，challenges and perspectives [J]. Molecular Psychiatry，2017，22：37-43.

[6] Hastie T，Tibshirani R，Friedman J. The Elements of Statistical Learning [M]. Berlin：Springer，2009.

[7] Luedtke A，Sadikova E，Kessler R C. Sample size requirements for multivariate models to predict between-patient difffferences in best treatments of major depressive disorder [J]. Clinical Psychological Science，2019，7：445-461.

[8] 生产Compass算法的公司Northpointe在一份报告中声称，对两组被告的暴力预测的准确性是相同的：大约70%的犯罪被正确预测（参见Dieterich等人的COMPAS风险量表：展示准确性公平性和预测性平等）。根据Northpointe的说法，黑人被告人和白人被告人之间不同程度的误报归因于他们犯罪流行率的不同基准率。可以让算法在具有不同基本率的组上获得相同级别的误报。然而，这是以降低准确性为代价的。有大量关于人工智能预测的公平性及其权衡的文献，关于这些算法的文本部分基于Cossins的研究。

[9] Angwin J，Larson J，Mattu S，et al. Machine Bias [M]. New York：ProPublica，2016.

[10] Hamilton M. The sexist algorithm [J]. Behavioral Sciences & the Law，2019，37：145-157.

往某市少数族裔人口比例较高的地区，无论这些地区的有效真实犯罪率如何[1]。此外，越来越多的用于执法的面部识别软件成为种族和性别偏见的另一个潜在来源[2]。另一个例子涉及亚马逊的“Rekognition”软件，该软件被一些警察部门和其他组织使用。2018年，美国公民自由联盟发现它错误地将国会议员与被指控犯罪的人进行匹配，将非裔美国人和拉丁裔国会议员误认为是照片中的人[3]。最近一项评估三种商业性别分类器准确性的研究表明，它们在对男性受试者进行分类方面的表现优于对女性受试者的分类，而所有这些分类器在肤色较深的女性中表现最差[4]。此外，最近的研究表明，如果不加以控制，词嵌入人工智能会表现出过时的性别刻板印象，例如“医生”是男性，“接待员”是女性[5]。

这些发现引发了关于使用人工智能进行风险评估的公平性的更广泛的辩论[6]。尽管算法风险评估可以被视为克服人类偏见的一种手段，但它们仍然可以反映成见和制度化的偏见。人工智能根据数据（如刑事档案）进行训练，这些数据本身可能反映了警察、检察官或法官的偏见。基于这些数据，该算法“得出”具有某些特征的群体比其他群体更危险，而实际上这是有偏见的数据的结果。这有时被称为“偏内偏外”。换言之，人工智能预测的结果高度依赖于所用数据的质量。使用神经影像数据而不是警察档案的一个优势可能是神经影像不能反映人类的偏见。人工智能寻找大脑活动和累犯之间的相关性。因此，人工智能神经预测可能提供减少风险评估偏差的可能性。然而，由于神经预测可能会被纳入现有的风险评估工具中[7]，只要一般算法中的偏见没有解决方案，偏见仍然是一个问题。

此外，风险评估有“典型的歧视性”[8]，因为它是根据群体特征将受试者分为低风险或高风险个体群体。累犯的神经标志物无疑在某些群体中比在其他群体中更为普遍。因为一个群体“大脑”的不同而以不同方式对待他们，会引发关于什么是不合理不平等待遇的难题。然而，这个问题并不是人工智能神经预测的典型问题，而是总体上风险评估和公平性的核心问题[9][10]。根据脑部扫描将人们分组，即使有助于防止可能的伤害，也很容易对那些被视为“高风险”的个人的生活的其他方面产生污名化和歧视性影响。根据大脑的形态来区分人可以成为一种现代颅相学。虽然某些制度程序可能会歧视那些被认为是“高风险”的人，但污名化可能是一个更具社会性的过程，会根据某些人的风险状况将其排除在外，例如，污名化可能是性犯罪者登

[1] Ensign D，Friedler S A，Neville S，et al. Runaway feedback loops in predictive policing [C] //Conference of Fairness，Accountability，and Transparency. Berlin，2018.

[2] Raji I，Buolamwini J. Actionable Auditing：Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products [C] //The 2019 AAAI/ACM Conference on AI，Ethics，and Society，2019.

[3] 参见https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

[4] Buolamwini J，Gebru T. Gender shades：intersectional accuracy disparities in commercial gender classification [C] // Conference on Fairness，Accountability and Transparency，2018.

[5] Bolukbasi T，Chang K W，Zou J Y，et al. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings [C] //Advances in Neural Information Processing Systems. Cambridge：MIT Press，2016，20：4349-4357.

[6] Berk R. Machine Learning Risk Assessments in Criminal Justice Settings [M]. Berlin：Springer，2018.

[7] Delfifin C，Krona H，Andine P，et al. Prediction of recidivism in a long-term follow-up of forensic psychiatric patients：incremental effffects of neuroimaging data [J]. PLoS One，2019，14：e0217127.

[8] Binns R. Fairness in machine learning：lessons from political [J]. Proceedings of Machine Learning Research，2017，81：1-11.

[9] Nadelhoffer T，Bibas S，Grafton S，et al. Neuroprediction，violence，and the law：setting the stage [J]. Neuroethics，2012，5：95.

[10] Tonry M. Legal and ethical issues in the prediction of recidivism [J]. Federal Sentencing Reporter，2014，26：167-176.

记的结果[1]。

其次，涉及隐私。用于预测累犯的神经数据和其他数据显然也可以用于其他目的。例如，保险公司评估客户或公司筛选求职者时，谁以及在什么条件下有权访问这些数据？保险公司是否可以访问，如果不可以，他们是否能够请求这样的程序来评估特定候选客户的风险？显然，在这种情况下，数据保护以及可能的访问是一个基本问题，在大数据时代中使用的算法已经引起了激烈的争论。目前关于同意性质的争论和公民对生物库中健康信息的控制程度之间也有相似之处。未来几年，关于遗传 / 健康信息和控制权（“生物权利”）商业化的讨论可能会加剧[2]。

第三，涉及负面的“自我实现预言”的可能性。这种疑虑来自最近的研究，研究表明接收遗传风险信息实际上可以影响接收人的行为、生理和主观体验，并改变他们的整体风险状况。斯坦福大学的研究人员发现，当人们被告知有肥胖或运动能力较低的遗传倾向时，获取这些信息会对他们的身体产生生理影响，改变他们对用餐或运动的反应。相关研究还发现对风险的看法改变了健康结果，因此那些被告知拥有高风险基因的人比那些被告知拥有保护性基因的人的结果更差[3]。根据这些发现，人们可能想知道，当你告知人们他们的风险信息时（无论是遗传的还是神经的），他们的心态会受到怎样的影响，以及这实际上如何改变他们的风险状况。这表明提供信息可能还需要道德和法律监管。

此外，仍不清楚如何将神经数据准确分类和概念化为风险因素。例如，在 Kiehl 等人（2018）的一项研究中，大脑年龄（灰质）的测量值用于预测累犯。实足年龄通常被认为是一个静态因素，但在提到大脑测量时，我们应该思考如何将它们概念化为风险因素。例如，考虑到大脑的可塑性，我们应该将大脑年龄视为动态风险变量还是静态风险变量？如果大脑年龄和正常年龄不同，我们如何评估罪犯，这将如何改变罪犯的神经预测特征？如果我们将神经数据视为动态因素，并且可以通过干预进行修改，那么我们可以谈论治疗目标和其他干预类型，而不是纯粹的“预测”。以这种方式使用神经预测可以通过更加个性化的惩教和社会康复措施来帮助预防犯罪，还可以使犯罪者更快地返回社区。如“个性化医疗”，它是一种使用个体的遗传和表观遗传信息来定制药物治疗或预防性护理的治疗方法[4]，神经预测有助于针对个体的“需求”进行干预。

目前，人工智能用于刑事司法系统，主要用于预测累犯。人工智能风险评估通常不提供犯罪的因果模型，因此其目的不是展示干预和降低风险的机会[5]。Barabas 等人（2018）得出当风险评估主要用作预测技术时，它们会助长大规模监禁和司法系统日益不平等的有害趋势的结论。

人工智能神经预测首先只是建立了大脑图像和累犯风险之间的相关性。然而，如果确实有可能开发基于神经数据的干预措施，这可能会为罪犯提供避免监禁的机会[6]。因为与无法改变的历史数据和其他风险变量（例如一个人的种族、年龄和性别等人口特征）不同，神经数据有可能成为新的康复干预和预防计划的目标，旨在减少接触精神病态特征的风险因素，并防止有风险的人在以后的生活中从事犯罪行为[7]。

---

［1］Tewksbury R. Collateral consequences of sex offffender registration［J］. Journal of Contemporary Criminal Justice，2005，21：67–81.

［2］Caulfifield T，Murdoch B. Genes，cells，and biobanks：yes，there’s still a consent problem［J］. PLoS Biology，2017，15：e2002654.

［3］Turnwald B P，Goyer J P，Boles D Z，et al. Learning one’s genetic risk changes physiology independent of actual genetic risk［J］. Nature Human Behaviour，2019，3：48–56.

［4］参见https://www.nature.com/subjects/personalized–medicine.

［5］Berk R. Machine Learning Risk Assessments in Criminal Justice Settings［M］. Berlin：Springer，2019.

［6］Nadelhof–fer T，Bibas S，Grafton S，et al. Neuroprediction，violence，and the law：setting the stage［J］. Neuroethics，2012，5：17–18.

［7］Ling S，Raine A. The neuroscience of psychopathy and forensic implications［J］. Psychology，Crime & Law ，2018，24：296–312.

这一点尤其重要，因为监狱环境可能对神经认知功能产生负面影响。事实上，研究发现监禁可能会导致自我控制能力下降[1]。尽管如此，干预的可能性也涉及其自身的道德和法律问题：对于犯罪者来说，可能很难在剥夺自由和接受（可能有些侵入性）治疗之间做出选择，尤其是考虑到拒绝医疗的权利[2]。然而，这又不是基于“人工智能神经预测”的干预的典型问题。

第四，涉及同意和强制。当这些技术得到充分开发并准备好使用时，可能会违反认知自由，迫使人们在未经同意的情况下进行扫描以用于量刑或惩罚[3][4]。胁迫，无论是技术上的还是道德上的抑或法律上的，不仅与所使用的武力有关，因为并非所有的成像技术都允许这样做，而且还与在无法拒绝的威胁或提议的背景下使用它们有关[5]。解决这个问题的一种方法是严格规范神经预测测试的知情同意。

第五，应注意神经影像学在法庭上施加的“诱人魅力”。陪审团和法官显然倾向于高估神经科学证据的准确性。尽管神经影像学旨在减少不确定性并提高法医环境的客观性，但由于证据评估中的认知偏差，在法庭上使用神经影像学存在误导的风险[6]。因此，引入神经预测可能会导致对神经数据的过度依赖。

此外，机器学习算法被认为是“决策黑盒”，其执行决策的方式利益相关者并不能完全理解，甚至专业数据科学家也不能完全理解[7]。我们必须谨慎对待所谓的“控制问题”，即人类操作员倾向于对机器自满、下放责任并过度依赖自主系统的输出，即使它们有偏见[8]。为了避免过度依赖，人工智能系统的透明度很重要，应向法官和陪审团解释它们是如何产生结果的[9]，应该使利益相关者能够适当地信任和管理这些工具，了解其在推理中如何给出特定输出以及基于什么理由。即使实际情况因大多数风险评估算法都是专有的这一事实而变得复杂，但对于社会来说，为了对其决策负责，人工智能算法可以被理解是非常重要的[10]。

值得注意的是，法律制度中可能有针对法庭上科学证据的可接受性的标准。例如，在美国法律环境中，Daubert 和 Frye 被用作标准。由于我们不关

[1] Meijers J，Harte J M，Meynen G，et al. Reduced self-control after 3 months of imprisonment：A pilot study [J]. Frontiers in Psychology，2018，9：69.

[2] Meynen G. Forensic psychiatry and neurolaw：description，developments and debates [J]. International Journal of Law and Psychiatry，2018，65：101345.

[3] Ligthart S. Coercive neuroimaging technologies in criminal law in Europe：exploring the implications for the prohibition of ill-treatment （article 3 ECHR） [M] //Regulating New Technologies In Uncertain Times Information Technology And Law Series. Berlin：Springer，2019：83-102.

[4] Meynen G. Ethical issues to consider before introducing neurotechnological thought apprehension in psychiatry [J]. AJOB Neuroscience，2019，10：5-14.

[5] Meynen G. Brain-based mind reading in forensic psychiatry：exploring possibilities and perils [J]. Journal of Law and the Biosciences，2017，4：311-329.

[6] Scarpazza C，Ferracuti S，Miolla A，et al. The charm of structural neuroimaging in insanity evaluations：guidelines to avoid misinterpretation of the fifindings [J]. Translational psychiatry，2018，8：227.

[7] London A J. Artifificial intelligence and black-box medical decisions：accuracy versus explainability [J]. The Hastings Center Report，2019，49：15-21.

[8] Pedreschi D，Giannotti F，Guidotti R，et al. Meaningful explanations of black box ai decision systems [C] //National Conference on Artificial Intelligence，2019，33：9780-9784.

[9] Gunning D，Aha D. DARPA's explainable artificial intelligence （XAI） program [J]. AI Magazine，2019，40：44-58.

[10] Weld S D，Bansal G. The challenge of crafting intelligible intelligence [J]. Communications of the ACM，2019，62：70-79.

注具体的法律制度，因此未对此进行更详细的讨论，但显然此类法律标准与法庭使用新技术有关[1]。

决定这些技术所需的准确性非常重要。当前的风险评估工具的 AUC 通常约为 0.70[2]。这对于此类算法是否足够或者阈值是否应该更高（如 0.80 或 0.90），都是在决定允许使用这种技术来预防犯罪之前必须做出的规范性选择。

此外，目前缺乏“真正的”预测模型。前文讨论的相关研究的一个局限性是它们不是谈论“纯”预测而是可以归类为后述研究，事后预测通常涉及基于事件发生后可用的信息对事件进行回顾性断言或推断[3]。但是，当应用于统计模型的背景下，预测和事后预测之间的区别在于对模型成功与否的评估是使用与建立模型相同的数据还是使用建立模型时未使用的新数据[4][5]。研究表明，用于预测应用的模型（如生物标志物）需要比标准统计方法更大的样本量[6]。此外，在之前讨论的研究中，有关累犯的神经标志物的数据是在犯罪后收集的，因此我们无法确定观察到的大脑差异何时出现[7]。未来的挑战是开发一个真正的预测模型，能够识别出犯罪风险最高的人，而神经影像学与人工智能相结合的研究可能是开发这种模型的关键。

最后，似乎还有一个更遥远的问题迫在眉睫。假设这些人工智能算法，无论是否有脑成像，都成为了很好的预测器，那会不会引入一种我们以前从未见过的确定性形式？人工智能系统可能被认为对将要发生的事情有一些“神圣”的预知，这可能会对人们体验和发挥的自由产生负面影响，对自由意志的信念似乎有积极影响[8][9]。

尽管如此，如今更紧迫的问题是我们并不擅长预测风险，即使是使用人工智能，我们仍然经常根据罪犯的假定危险性来实施制裁。如果人工智能在神经影像学的帮助下变得更加准确，它可以减少被错误归类为高风险的人数，从而减少实际上不合法的制裁，有助于中断所谓的“犯罪循环”[10]。

## 七、结论

要在刑事司法系统中实施结合神经科学和基于人工智能的暴力风险评估工具，还需要更加深入的研究。尽管如此，人工智能已经被用于刑事司法系统。由于这类技术的深远影响，以及近年来的快速发展，考虑道德和法律问题非常重要。除了讨论预测分析的技术限制和陷阱外，我们还确定了六个值得关注的关键问题：处理偏见、隐私、“自我实现

[1] Shats K，Brindley T，Giordano J．Don’t ask a neuroscientist about phases of the moon：applying appropriate evidence law to the use of neuroscience in the courtroom [J]．Cambridge Quarterly of Healthcare Ethics，2016，25：712-725.

[2] Douglas T，Pugh J，Singh I，et al．Risk assessment tools in criminal justice and forensic psychiatry：the need for better data [J]．European Psychiatry，2017，42：134-137.

[3] Yamada Y，Kawabe T，Miyazaki M．Awareness shaping or shaped by prediction and postdiction：editorial [J]．Frontiers in Psychology，2015，6：166.

[4] Gauch H G，Zobel R W．Predictive and postdictive success of statistical analyses of yield trials [J]．Theoretical and Applied Genetics，1998，76：1-10.

[5] Hastie T，Tibshirani R，Friedman J．The Elements of Statistical Learning [M]．Berlin：Springer，2009.

[6] Varoquaux G．Cross-validation failure：small sample sizes lead to large error bars [J]．Neuroimage，2018，180（A）：68.

[7] Cope L M，Ermer E，Gaudet L M，et al．Abnormal brain structure in youth who commit homicide [J]．NeuroImage：Clinical，2014，4：800-807.

[8] Crescioni A W，Baumeister R F，Ainsworth S E，et al．Subjective correlates and consequences of belief in free will [J]．Philosophical Psychology，2016，29：41-63.

[9] Feldman G，Chandrashekar S P，Wong K F E．The freedom to excel：belief in free will predicts better academic performance [J]．Personality & Individual Differences，2016，90：377-383.

[10] Barabas C，Virza M，Dinakar K，et al．Interventions over predictions：reframing the ethical debate for actuarial risk assessment [C]//Conference on Fairness，Accountability and Transparency，2018：62-76.

预言”的可能性、强制和同意、神经影像数据的吸引力以及对可解释的人工智能系统的需求。最后，我们指出了一个更遥远的问题，即高度准确的预测如何引入一种我们以前从未见过的确定性形式，但这仍然很遥远。

尽管如此，我们还是要强调，出于安全和正义的原因，准确的风险预测非常有价值。因此，原则上，我们认为至少应该探索可能在这方面有所帮助的技术，并在准备充分时将其用于刑事司法和法医精神病学领域。此外，神经预测和人工智能在某种程度上带来了新的伦理和法律挑战，我们应在使用这些技术之前处理好这些挑战。更具体地说，我们必须找到解决方案来防止系统反映人类偏见，以使其能够提供客观和值得信赖的数据。

因此，我们认为，在刑事司法和法医精神病学中使用基于人工智能的系统应受到实质性监管，以保护公民免受系统错误或滥用的影响。在此基础上，不仅在这些技术完全可用时，而且在它们处于研究和开发阶段时，我们都强调准确的危害或益处分析的重要性。

（责任编辑：何　为）

# Neuroprediction and A.I. in Forensic Psychiatry and Criminal Justice: A Neurolaw Perspective

Leda Tortora[1] Gerben Meynen[2,3] Johannes Bijlsma[2] Enrico Tronci[4] Stefano Ferracuti[1]
Wan Huan (Translator)

*1. Department of Human Neuroscience, Sapienza University of Rome, Rome;*
*2. Willem Pompe Institute for Criminal Law and Criminology/Utrecht Centre for Accountability and Liability Law (UCALL), Utrecht University, Utrecht;*
*3. Faculty of Humanities, Vrije Universiteit Amsterdam, Amsterdam;*
*4. Department of Computer Science, Sapienza University of Rome, Rome*

**Abstract:** Advances in the use of neuroimaging in combination with A.I., and specifically the use of machine learning techniques, have led to the development of brain-reading technologies which, in the nearby future, could have many applications, such as lie detection, neuromarketing or brain-computer interfaces. Some of these could, in principle, also be used in forensic psychiatry. The application of these methods in forensic psychiatry could, for instance, be helpful to increase the accuracy of risk assessment and to identify possible interventions. This technique could be referred to as “A.I. neuroprediction” and involves identifying potential neurocognitive markers for the prediction of recidivism. However, the future implications of this technique and the role of neuroscience and A.I. in violence risk assessment remain to be established. In this paper, we review and analyze the literature concerning the use of brain-reading A.I. for neuroprediction of violence and rearrest to identify possibilities and challenges in the future use of these techniques in the fields of forensic psychiatry and criminal justice, considering legal implications and ethical issues. The analysis suggests that additional research is required on A.I. neuroprediction techniques, and there is still a great need to understand how they can be implemented in risk assessment in the field of forensic psychiatry. Besides the alluring potential of A.I. neuroprediction, we argue that its use in criminal justice and forensic psychiatry should be subjected to thorough harms/benefits analyses not only when these technologies will be fully available, but also while they are being researched and developed.

**Key words:** Neuroprediction; Artificiaintelligence; Recidivism; Forensicpsychiatry; Riskassessment; Neurolaw