刑事诉讼中预测性算法应用的权利风险及防范

孙法柏　刘清华
(山东科技大学 文法学院,山东 青岛 266590)

[摘　要] 在数字时代背景下,预测性算法逐步应用于刑事诉讼领域,其所伴生的权利风险亟需应对。预测性算法的不当使用存在人权侵害的现实可能性,难以契合《世界人权宣言》《公民权利及政治权利国际公约》等国际人权法的基本精神。具体而言,预测性算法易导致法官基于先入为主的风险预判作出不公正裁决;控辩格局因控辩双方在算法资源与能力上的不对称而失衡,被告权利难以保障;算法运行机制与无罪推定原则之间存在内在张力,自动化决策亦可能削弱被告的知情权与诉讼参与权。为防范上述权利风险,应合理规范预测性算法应用,坚守司法公正与权利保障底线;通过严格设定预测性算法准入条件,建立算法解释与思考路径追踪机制,并构建以审判为中心的控辩平衡制度,始终把预测性算法限定于辅助性、从属性的地位。

[关键词] 刑事诉讼;预测性算法;程序正义;风险防范
[中图分类号] D925.2　[文献标识码] A　[文章编号] 1008-7699(2026)01-0037-12

一、问题的提出

司法界在20世纪中后期开始尝试将AI引入程序,相继开发了法律推理、案件模拟分析等司法AI系统。随着循证方法(Evidence-based Approach)的引入,预测性算法逐渐成为刑事诉讼的重要工具。① 该算法基于历史数据对输入变量进行权重计算,可以对被告不出庭、再犯及实施其他不当行为等风险进行预估,②辅助司法人员作出相关决策。然而,预测性算法在提升司法决策科学性和效率性的同时,其合法性、公正性在实践中也颇受质疑,亟待审慎审查与规范。

预测性算法的运行机理与刑法的特殊预防理论具有内在契合性。③ “选择性失能”(Selective Incapacitation)理论主张,相较于初犯、偶犯,惯犯应对多数犯罪承担主要责任。因此,通过对惯犯实施针对性控制措施,可以有效降低整体犯罪率。④ 该理论突破了传统刑法强调个体本位与被告人权利保障的既定范式,推动了刑法理念向风险治理与社会防卫取向的深刻转型。其理念更新以维护整体安全与遏制不稳定因素扩散为目标,试图缓解因犯罪率上升而引发的社会恐慌情绪,⑤从而重塑了刑法介入的正当性。

无论从历史发展还是理论逻辑来看,刑事司法体系引入预测性算法都是一种必然趋势。然而,

[收稿日期] 2025-10-22
[基金项目] 山东省高级人民法院2025年度重点调研课题(SDGYKT2025A05-2)
[作者简介] 孙法柏(1970—),男,山东新泰人,山东科技大学文法学院教授,博士。

① 楼伯坤、鲍博:《刑事司法中风险评估的算法风险与规制措施》,《学术交流》2024年第3期,第62页。
② Megan Stevenson, *Assessing Risk Assessment in Action*, Minnesota Law Review, Vol. 103, No. 1, 2018, pp. 303-384.
③ 江溯:《自动化决策、刑事司法与算法规制——由卢米斯案引发的思考》,《东方法学》2020年第3期,第77页。
④ Bernard E. Harcourt, *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*, University of Chicago Press, 2007, p. 88.
⑤ 参见前注①,楼伯坤、鲍博文,第66页。

欧美(特别是美国)学界围绕该议题的研究与争议从未停止:一方面,部分学者秉持积极接受科技变革的态度,倡导在刑事司法各环节广泛采用算法工具,以提升治理效能;另一方面,更多学者则从人权保障和程序正义的视角出发,对预测性算法的合理性、透明度与正当性提出质疑与批判。①

卢米斯案(State v. Loomis)可谓是有关预测性算法的里程碑案件,2016 年 7 月,威斯康星州最高法院作出维持原判的终审裁决,引发广泛争议。② 该案凸显出,预测性算法的介入将直接影响正当程序所强调的被告刑事诉讼权利。由于这些权利已被普遍视为现代人权体系的重要组成部分,因此需在国际人权法框架下审视预测性算法的使用边界与规范要求。

二、预测性算法的人权"限制"的规范与法理分析

在刑事诉讼程序中使用预测性算法,已在众多著述以及司法实践中被指出可能对被告的基本权利产生不利影响。算法介入刑事司法是否构成对被告基本权利的正当限制,国家是否在合法框架内对既有人权保障范围作出收缩,构成了分析该不利影响是否侵害人权的前提。唯有厘清这一基础问题,方能对算法适用的正当性、边界与规范路径进行探讨。

(一)"依法限制"的规范要求

由于实在法中的人权规定本身即蕴含国家权力与个人权利之间的对立统一关系,因此必然承认在特定情境下对人权进行限制的可能性与必要性。

《世界人权宣言》(以下简称《宣言》)首次对人权基本内涵作出系统规定,构成后续人权条约与众多国家宪法的理论基础。③ 其在序言末尾明确指出:"……通过国家的和国际的渐进措施,使这些权利和自由……得到普遍和有效的承认和遵行。"《宣言》全文内容基本为对公民的赋权性规定,并未就人权限制事宜具体说明,仅在第 29 条第 2 款规定:"人人在行使他的权利和自由时,只应受法律所规定的限制,其目的……",《宣言》认为对人权的限制应"受法律所规定",且根本目的在于保障他人合法权益或符合社会道德、秩序与福祉。

由于《宣言》主要以国家作为保障人权的义务主体,故"受法律所规定"这一要求中的"法律"应指各国国内法。也即,人权限制事项应由各国国内法明确规定。因此,若预测性算法构成对人权的限制,则应由国内法将其明确规定为具有"合法目的"的合理限制。

欧美、日韩等国均已出台有关刑事司法或司法领域人工智能的规制文件(含法律、指南、规则等);④我国尚未针对司法领域 AI 应用进行专门立法,但通过最高人民法院相关意见及人民检察院量刑建议指导意见,⑤支持人工智能在案件办理全流程及量刑辅助等环节的应用。从若干代表性国家和地区的最新规定来看,除欧盟与中国外,大多数司法辖区尚未出台涉及规范刑事领域预测性算法的法律或规范性文件,而多依赖行业自律或针对 AI 的一般性原则规定。这一状况表明,许多

① 如 Angwin et al. 2013,针对算法的种族偏见;De Miguel Beriain 2018, Dressel & Farid 2018,针对预测性算法的准确性;Freeman 2016,针对算法的私人属性。

② State v. Loomis. 88I N. W 2d 749 (Wis. 2016). 终审法院认为:第一,COMPAS 所获取的数据具有公开性,卢米斯作为被告有权否认或解释据以形成结果的相关数据以保障其准确性;第二,法院仍享有自由裁量权,并非仅基于该报告作出判决,因而若只是将 COMPAS 评估结果作为参考,并未侵犯被告人获得个性化判决的权利;第三,风险评估以性别为因素能提高预测精确性,而非出于歧视目的,卢米斯无足够证据证明原审法官实际考虑了性别因素。

③ 罗艳华:《〈世界人权宣言〉:全球人权治理的重要基石》,《中国国际战略评论》2018 年第 1 期,第 209-210 页。

④ 如欧盟《人工智能法》及其委员会颁布的《关于禁止人工智能系统实践的指南》(草案);美国加州司法委员会发布的规则 10.430 与标准 10.80;英国最新修订的《英国司法人员使用人工智能指南》;韩国的《人工智能发展及信任基础建立基本法》以及日本的《人工智能相关技术研究开发及应用推进法》等。

⑤ 如我国《最高人民法院关于规范和加强人工智能司法应用的意见》以及《人民检察院办理认罪认罚案件开展量刑建议工作的指导意见》。

国家尚不具备构建相关完整规范体系的现实条件，只能在技术发展先于立法的背景下，由司法实践率先摸索相对可行的路径。事实上，将预测性算法引入刑事司法程序符合技术革新的内在趋势，具备一定的现实合理性，但这种合理性并不必然等同于法理上的正当性。尤其在缺乏明确法律规制的前提下，其应用难以满足《宣言》所要求的“依法限制”标准。

以美国为例，其每年耗费1820亿美元以维持本国约200万人口的监禁体量。① 为控制居高不下的监禁率，2002年，弗吉尼亚州开始在全州范围内对非暴力重罪罪犯使用风险评估工具。② 肯塔基州于2011年通过HB 463法案(“House Bill 463”(HB 463) of Kentucky)③，默许法官在审判前使用风险评估算法以降低监禁率。抑制监禁率是美国法院在刑事诉讼中引入预测性算法的重要原因之一，降低监禁率可以大幅节约财政开支、释放司法资源并提高司法系统效率，该举措似乎最接近于促进社会一般福利的“人权限制”目的。

然而，AI虽可提高效率，但法律的向度应更倾向于公正。④ 预测性算法作为技术工具，确可在数字时代为法官提供信息支持、提升裁判效率，但其正当功能应止于辅助决策，而不得转化为加重羁押、量刑或监管等刑事不利后果的依据。在未穷尽非刑罚性替代措施、亦未证明既有制度不足以应对相关风险之前，仅凭算法预测结果对被告科以更为严苛的处罚，实质上是将未经证实的未来危险性提前转化为现实惩罚，这不仅违反人权限制所要求的必要性与最小侵害原则，也偏离了“依法限制”应服务于正当目的的规范要求。

从国际人权法角度看，《公民权利及政治权利国际公约》(以下简称《公约》)明确将国家定位为人权的促进与保障者，其所允许的消极性人权限制仅存在于第4条规定的公共紧急状态之下，且须严格符合紧急性、必要性、禁止滥用与非歧视等条件。刑事诉讼引入预测性算法既不以公共紧急状态为前提，也未证明在穷尽替代手段后其仍属不可或缺，因而难以满足《公约》所设定的人权限制标准。《公约》并不能为预测性算法在刑事司法中的适用，尤其是据此加重被告的刑事不利后果，提供规范上的正当性依据。

(二)人权限制的法理分析

对人权的限制，有学者基于“权利与义务相统一”的思路，主张个人在享有权利的同时应当承担对社会整体的责任与义务，从而接受针对权利的限制；《公约》若干条款明确了人权内容并对其进行范围界定，因此《公约》亦体现了权利与义务相统一的精神。⑤ 该学者的理解可能构成对《公约》及人权的误读：其一，该观点混淆了个体的社会责任与义务同国家对人权施加的限制；其二，其逻辑预设人权的享有以履行社会责任为前提，实质上否认了人权的无条件性与普遍性。

正如前文所述，《公约》明确将国家而非个人定位为人权保障的义务主体，不能将“权利与义务

① United States of America | World Prison Brief (prisonstudies.org), at https://www.prisonstudies.org/country/united-states-america; Wendy Sawyer & Peter Wagner, *Mass Incarceration: The Whole Pie* 2025, at https://www.prisonpolicy.org/reports/pie2025.html (Last visited on December 31, 2025).

② 其量刑委员会的一份报告显示，该工具的使用将接近一半的被告以监狱以外的替代方案进行安置。详见Julia Angwin et al., *Machine Bias*, at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Last visited on October 27, 2025).

③ 该法案也被称为“The Public Safety and Offender Accountability Act”，详见2011 Bill Text KY H.B. 463- CSG South.

④ 《“美国法律体系中的人工智能：挑战与问题”讲座顺利举行》，中国政法大学法学院，https://fxy.cupl.edu.cn/info/1087/6473.htm，2025年11月12日访问。

⑤ 朱晓青、柳华文：《〈公民权利和政治权利国际公约〉及其实施机制》，中国社会科学出版社2003年版，第27页。

相统一”片面理解为权利与义务须共存于同一主体。尽管《公约》对个体的社会责任与义务有所提及,[①]但该责任与义务仅具有伦理与道德意义,并不构成个人享有人权的法律前提,更不能作为限制人权的正当依据。

唐纳利曾举例,国家可因某人发表恶意损害他人的言论而对其施以惩罚,但依据在于受害人被侵犯的合法权益,而非行为人的表达自由。[②] 劳特派特(H. Lauterpacht)亦认为个体自然权利的必要界限存在于他人的自然权利中。[③] 也即,个体承担权利限制的依据,在于他人或社会将因其超越限制的行为蒙受损害。只有在他人或社会的权益事实上面临威胁,且此种威胁与个体行使权利的行为具有因果关系时,限制个体人权才具备相当的条件。然而,预测性算法的结论并非准确映射现实,而是由数字逻辑所构建的概率化推测。这种推测并不代表他人或社会的权益遭受侵害的必然事实,在个体尚未作出行为时,也无从谈起因果关系。行为人应仅就已犯之罪承担刑责,而无需就其罪责以外、其所未实施的行为承担额外的不利后果。因此,仅依据尚未发生的可能性便限制相关人员的权利,实在难言合理正当。

个人或群体可能因其行为对他人权利造成不利影响而无法全面行使一些人权,但这一问题在概念上有别于这些人权对于个人或群体的自始可适用性。[④] 被限制的人权已不再等同于人们最开始所享有的人权。基于人权的普遍性与优先性,对人权的限制只能作为非常态的例外措施,而不得在常态时期被制度化、工具化地反复适用。正因如此,无论《宣言》还是《公约》,均对人权限制设定了极为严格的条件。与此相对,作为技术革新的自然产物,预测性算法在刑事司法中的运用具有持续性与常设性,其对权利的影响亦随之长期存在,这显然难以契合人权限制所要求的特殊性与临时性。

综上,无论从国际人权规范,还是从人权理论出发,在刑事诉讼中引入预测性算法的做法均难以通过人权限制理论获得解释与正当性。因此,从国家义务的角度看,国家并无正当理由放任被告承受预测性算法对其权利造成的持续性不利影响。

三、程序正义视角下刑事被告权利的算法挑战

既然刑事诉讼引入预测性算法难以通过人权“正当限制”理论获得解释,其对被告权利的影响更应被理解为对基本权利的侵害。预测性算法通过嵌入与程序正义密切相关的制度环节,对被告在诉讼中的权利产生深远影响。

(一)算法渗透对审判独立性的削弱

《公约》第14条第1款就审判中立与审判独立原则作出规定,[⑤]《中华人民共和国刑事诉讼法》(以下简称《刑事诉讼法》)第5、6条亦分别就司法机关依法独立行使职权与保障司法公正作出规

① 《公约》前言载有“明认个人对他人及对其隶属之社会,负有义务,故职责所在,必须力求本盟约所确认各种权利之促进及遵守……”。

② Jack Donnelly, *Human Rights and Asian Values: A Defense of “Western” Universalism*, Cambridge University Press, 1999, p. 79.

③ H. Lauterpacht, *International Law and Human Rights*, Stevens and Sons, 1950, reprinted in 1968, p. 366.

④ René Provost, *Reciprocity in Human Rights and Humanitarian Law*, British Yearbook of International Law, Vol. 65, No. 1, 1994, pp. 383-454.

⑤ 即“人人在法院或法庭之前,悉属平等……应有权受独立无私之法定管辖法庭公正公开审问。”

定。[①] 审判的独立与公正不仅旨在构建良好的诉讼环境，本身亦蕴含着保障被告程序权利的重要法律价值。鉴于刑事诉讼法以保障人权为基本目的，其制度设计皆将保护人权纳入考量，因而可将审判中立与审判独立视为在宏观层面统摄并支撑程序权利保障的基础性制度。

在刑事司法与预测性算法深度融合的背景下，算法的技术逻辑与运行特性正逐渐嵌入刑事诉讼体系，对司法实践产生持续影响，其中以对法官裁判理念与判断方式的重塑最为显著。亟需检视这种影响是否危及审判中立与独立的制度根基，进而侵蚀了被告的程序性权利。

首先，预测性算法通过整合并分析海量信息与数据，对法官裁判形成持续性的“指导”甚至“预判”，这在一定程度上诱发了决策让渡，本质是算法权力利用技术优势瓜分法官审判权。[②] 2023 年 1 月 30 日在“全球 AI 审判第一案”中，哥伦比亚法官 Juan Manuel Padilla García 将 ChatGPT 有关法律问题的部分回答直接写入判决；[③]2023 年 11 月 13 日，巴西北部阿克里州的 Jefferson Rodrigues 法官使用 ChatGPT 撰写了一份充满细节错误的判决书……[④]。这些例证均表明算法结论已在不同程度上介入并影响司法判断过程。这反映出一种潜在的法律威胁，即法官过度信任算法所呈现的“客观性”与“科学性”，弱化甚至替代了自身的独立判断。[⑤]

在量刑环节中，预测性算法所引发的这一“锚定效应”亦值得警惕。法官在刑事诉讼中过度倚重技术，致使他们难以审慎处理案件，难以意识到这对被告人产生的影响，最终导致对被告命运的漠视。法官不应简单以被告的风险评估分数为主要依据，更不能据此施加更严苛的刑罚。遗憾的是，即使没有任何现实证据证明被告会再犯，算法结果也会削弱法官的中立性与内心确信。简言之，法官在数据面前极易形成先入为主的风险认知，进而可能高估被告的社会危害性，并在量刑或假释裁量中不自觉地趋于从严处理。

综上，预测性算法在刑事司法中的应用，凭借其技术权威性与数据外观上的客观性，易对法官裁判产生“锚定效应”，弱化其独立判断并植入潜在偏见，从而使裁判权面临被技术逻辑主导甚至替代的风险。在此过程中，算法不仅可能导致刑罚结果不当加重，也使被告在程序中难以有效对抗既定的风险预判，进而实质侵蚀公正审判原则。在预测性算法逐步嵌入司法运行体系的现实背景下，若缺乏明确有效的规范约束，其对被告程序权利的持续性影响将持续扩张。因此，如何在承认技术辅助价值的同时，通过制度设计防止算法“越权”、维护裁判独立与程序正义，构成刑事司法数字化进程中亟需回应的核心问题。

（二）预测性算法对被告程序权利的侵蚀风险

算法通过削弱审判中立性与独立性，动摇程序正义的制度基础，间接侵害了被告的刑事程序权利。在此基础上，算法还通过诸多方式直接侵害被告理应享有的其他程序权利。

---

① 《刑事诉讼法》第 5 条规定“人民法院依照法律规定独立行使审判权，人民检察院依照法律规定独立行使检察权，不受行政机关、社会团体和个人的干涉。”第 6 条规定“人民法院、人民检察院和公安机关进行刑事诉讼，必须依靠群众，必须以事实为根据，以法律为准绳。对于一切公民，在适用法律上一律平等，在法律面前，不允许有任何特权。”

② 李训虎：《刑事司法人工智能的包容性规制》，《中国社会科学》2021 年第 2 期，第 51 页。

③ Blu Radio, *Sentencia La Tomé Yo, ChatGPT Respaldó Argumentación: Juez de Cartagena Usó Inteligencia Artificial*, at https://www.bluradio.com/judicial/sentencia-la-tome-yo-chatgpt-respaldo-argumentacion-juez-de-cartagena-uso-inteligencia-artificial-pr30 (Last visited on January 21, 2026).

④ Agence France-Presse, *Brazil Judge Under Probe for AI Errors in Ruling*, at https://www.abs-cbn.com/overseas/11/14/23/brazil-judge-under-probe-for-ai-errors-in-ruling (Last visited on December 31, 2025).

⑤ Elizabeth E. Joh, *Policing by Numbers: Big Data and the Fourth Amendment*, Washington Law Review, Vol. 89, No. 1, 2014, pp. 35-68.

1.控诉权联合算力加重诉讼构造失衡:控辩平等的程序性危机

公权同人权是一对矛盾,公权因保护人权而生,但也天然拥有压制人权的倾向。正是在这一结构性矛盾中,控辩平等原则成为平等理念在国家与个人的刑事诉讼关系中的集中体现。① 该原则要求实现刑事被告与控诉机关的平等,旨在确保控诉机关与被告的法律地位与程序权利处于对等状态,从而防止国家权力在刑事追诉中取得压倒性优势。

从功利角度而言,刑事诉讼作为国家化解冲突、维持公共秩序的重要制度安排,其正当性并不在于单向度追求惩罚效率,而在于以相对低成本、理性且可接受的方式解决争议,进而缓解社会紧张、防止不稳定因素扩散。哈贝马斯曾言,只有沟通才能求取真正的正义,这需要人们通过理性对话表达出正当的合意。② 这种经充分交涉的"合意",只有在控辩双方处于实质平等的程序地位时才可实现。

当下,民众生活选项处处受制于算法筛选,企业、机关等主体因掌握数字能力而更为强大。③映射至刑事诉讼领域,控辩双方的差距将进一步被技术拉大。实践中,在资源方面占优的控、审机关接受AI服务,可以掌控海量司法数据,从而成为数据优势主体;控方甚至可以依法行使国家权力要求商业数字平台等社会力量提供相关数据协助司法。而相对作为"数字客体"的被告方则变得弱势,仅能行使一定程度的数据访问权利;各国普遍确立的侦查秘密原则还进一步提高了被告方获取相关资料的门槛。

由此,控辩双方之间形成难以逾越的"数字鸿沟"。④ 这样的现实不仅导致刑诉构造进一步失衡,也将限制被告调查取证能力,并深刻影响其在后续审判程序中有效行使辩护权。相比较算法对审判者行权思维的间接塑造,控方接受算法"加权"的现实将会对被告产生更直接、紧迫的危险。

根据尊严价值理论,法律程序是否正当应以对人之尊严的维护程度作为评价标准,也即验证法律程序内核是否以人为中心、是否以人性为基础。⑤ 然而,在披着国家侦控职能合法外衣的预测性算法实践中,对被告个人数据的采集与处理绕过了征得其同意的步骤。由于风险评估成为司法权运行的一部分,可以说,为预测性算法提供充足数据以支持其结论,并在此基础上接受国家的审判,是被告的责任和义务,这一现象在采用职权主义诉讼模式的国家中尤为突出。在这种情况下,被告丧失部分诉讼主体地位,成为了司法体系中的数字客体,⑥也进一步放大了预测性算法对被告程序权利的侵蚀风险。

总体而言,预测性算法嵌入刑事司法运行后,被告受到了技术赋权的国家追诉体系与算法机制的双重作用,其诉讼主体性面临弱化风险,进而影响其诉讼参与权的享有与程序平等的实质实现。⑦ 在此背景下,尽管我国现行刑事诉讼法体系已确立多项控辩平等保障机制,相关制度仍有必

① "控辩平等原则"亦被域外刑事诉讼学界称为"手段同等原则",其意为原则上应当以对待控诉机关一样,平等对待被告。详见[德]约·阿希姆·赫尔曼:《德国刑事诉讼法典》中译本引言,李昌珂译,《德国刑事诉讼法典》,中国政法大学出版社1995年版。

② 冯颜利、张朋光:《哈贝马斯的正义观与当代价值——兼论哈贝马斯与罗尔斯正义观的主要异同》,《华中师范大学学报(社会科学版)》2013年第6期,第57-60页。

③ 於兴中:《算法社会与人的秉性》,《中国法律评论》2018年第2期,第59页。

④ 即使有的辩护人及其所属单位购买了人工智能服务,这项成本最终也会分摊至被追诉人身上。详见谢澍:《人工智能如何"无偏见"地助力刑事司法——由"证据指引"转向"证明辅助"》,《法律科学》2020年第5期,第114页。

⑤ 陈瑞华:《程序正义的理论基础——评马修的"尊严价值理论"》,《中国法学》2000年第3期,第145-146页。

⑥ Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System*, U. C. Davis Law Review, Vol. 52, No. 2, 2018, pp. 1067-1118.

⑦ 司法民主的保障需要当事人有效行使诉讼参与权,而诉讼参与权的观念基础为诉讼主体性理念。

要被针对性完善，以避免控辩平等原则在数字化条件下实质空转。

2. 预测性算法与无罪推定原则的法律张力

现代刑事诉讼立法普遍确立了无罪推定原则，[①]此为法律弥补控辩双方力量差距的重要举措。无罪推定原则可作为证据规则用于分配证明责任：由控方举证被告有罪，而后者无需承担证明自身无罪或罪轻的责任；若控方无证据或证据不足以证明其控诉内容，则推定被告无罪。[②]

部分预测性算法涉及相当比例的心理调查类设问，此类工具的定位更接近于心理健康评估工具。[③] 在严谨的法律程序中，将心理评估工具用以预测被告的人身危险性是功能错位的行为。此外，预测性算法虽围绕被告设问，但十分关注被告所在社群的环境因素。然而，生活在贫困环境或受教育程度较低，并不意味着被告一定会再次犯罪，这些因素只能表明一种统计学范畴的相关关系，而非客观的因果关系。因此，风险评分不一定能揭示被告是否具有危险性，[④]但其结论本身确实预设了被告未来有罪的情形。

无罪推定原则还要求司法机关在证据存疑时，应作出有利于被告的裁决。然而，多数预测性算法为商业公司私有，其开发理念为利益最大化思维，这注定了算法设计缺乏对人权价值的重视。例如，当算法误判被告有较高再犯风险，并建议对其人身自由施加更严格的限制时，这种错误往往不易被察觉；相反，若算法误判被告再犯风险较低，并建议对其施加更少限制，那么一旦被告再次犯罪，这种错误便会引起公众关注。后者无疑会对算法开发公司的声誉造成负面影响，因此，公司很可能倾向于预防后者而忽略前者，[⑤]这样的倾向显然与“疑罪从无”“疑罪从轻”的精神背道而驰。

因此，若将预测性算法视为刑事诉讼程序组成部分，甚至将其自身运行过程视为一种“微观刑事诉讼程序”，算法机制与无罪推定原则之间便不可避免地产生张力。《宣言》在第 11 条第 1 款明确了无罪推定原则，我国《刑事诉讼法》第 12 条与第 51 条亦蕴含了无罪推定精神。然而，预测性算法所引发的正当性危机至今仍未在技术层面与理论层面获得妥善回应。其以心理健康调查为定位、以相关关系为决策标准得出的被告再犯结论，在法理上难言未抵触无罪推定原则的精神。

3. 算法结论的优绩主义倾向：诉讼参与性与公开性的淡化[⑥]

奥克塔维奥·特耶罗（Octavio Tejeiro）曾提及，人们对 AI 替代法官的忧虑容易导致法律层面的道德恐慌。尽管如此，他依然认为这样的工具会被迅速接受和普及。[⑦] 的确，人类法官基于自身经验、价值判断与情感等因素作出裁决，不可避免会陷入个体局限性中，难以从全方面完整视角对案件进行研判；而 AI 具备极致的数据存储量和分析能力，能够在效率与一致性层面呈现出人类裁判者难以企及的“优绩”优势。

2025 年 9 月 15 日，“Justice-1”人工智能系统在加拿大温哥华市法院庭审现场仅仅耗时 12 秒

① 无罪推定原则可追溯至 1764 年贝卡利亚的《论犯罪与刑罚》：在法官判决之前任何人不得被视为有罪，只要还未确定其已侵犯了有关公共保护的契约，社会就不得免去对他的公共保护。贝卡利亚的论述是无罪推定原则最广为熟知的含义。

② James Bradley Thayer, *The Presumption of Innocence in Criminal Cases*, Yale Law Journal, Vol. 6, No. 4, 1897, pp. 185-212.

③ 张振声：《犯罪人风险行为评估技术新进展——COMPAS 系统评介》，《辽宁公安司法管理干部学院学报》2022 年第 3 期，第 5 页。

④ Julia Angwin et al., *Machine Bias*, at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Last visited on October 27, 2025).

⑤ 江溯：《自动化决策、刑事司法与算法规制——由卢米斯案引发的思考》，《东方法学》2020 年第 3 期，第 79 页。

⑥ 鉴于文章目的在于揭示预测性算法对被告程序权利的影响，本文所讨论的诉讼参与性与诉讼公开性较为狭义。

⑦ Deena Theresa, *Colombian Judge Uses ChatGPT in Ruling, Makes Humane Decision*, at https://interestingengineering.com/innovation/chatgpt-makes-humane-decision-columbia (Last visited on January 15, 2026).

便完成了对当事人陈辞、证据与法条的智能交叉比对,并自动生成长达83页的裁判文书。① 在我国,AI更是实现了3秒生成90%完成率的判决书、10分钟完成2日人力工作量的类案检索、单日辅助审结16件案件等成就。② 不可否认,人工智能已在多个领域远超人类能力。此时的法律从业者必须更强烈地意识到,求真与效率仅是法律价值的一隅,它们远不能实现法律的所有意义。算法在最终结论的产生层面或许能比肩业务熟稔的资深法律从业者,但在法律领域,完美结果的达致从不取决于结果本身。

依据人民主权理念,作为权力的最终来源主体,民众理所当然参与到国家事务以及涉及自身利益的决策中。此外,"程序的价值在于以看得见的方式实现正义"这句话表明,正义是否实现、正义的实现方式是否被呈现,都需要"得到他者的认可",这种他者视角便是民众视角,这构成了司法民主化的基础。鉴于此,刑事诉讼程序的设置应当强调当事人与公众对诉讼的参与性。其中,保障当事人诉讼参与权的最终目的,是确保其可以对涉及自身利益的裁判施加最大程度影响。然而,这一切的前提在于当事人的知情权被充分尊重,诉讼公开性则为知情权的享有提供条件。

司法人员是否客观中立、是否尊重当事人,以及当事人参与诉讼的程度都是评判司法程序的指标。③ 程序的设置是否符合公众对正义的期待,很大程度上取决于程序本身是否满足了当事人"被看见"的需求。当事人观点被其他诉讼主体所知悉并考虑,这能使其在程序中切实相信自己对结果的产生发挥了作用。因此,程序的构建应使当事人充分感知其与决策者之间的联系,并足以让当事人意识到决策者在与自己交涉时能够真正共情于自己的感受。正因如此,各国刑事诉讼法普遍强调法官的亲历性与被告人辩护权等制度,并在程序全过程嵌入保障当事人参与权的民主法治精神。然而,人工智能司法尤其是预测性算法的出现打破了人们长久以来所形成的程序平衡。

首先,预测性算法动摇了当事人参与刑事诉讼的心理基础。根据相似性吸引原则(Principle of Similarity Attraction),个体更倾向于喜欢在信念、价值观、态度、个性特征等方面与自己相似的人;同时,证实性偏见(Confirmation Bias)解释了人们为何更相信他们愿意相信的事物,而对未迎合其喜好的客观真相视而不见。在刑事诉讼程序中,若当事人认为决策者与自己共享经验常识与普遍理性,且坚信决策者具备与自己相似的情感,他们便可能通过庭审实现与决策者的亲身交流,建立起对法院裁判的道德权威的信任。④ 但AI不具备第一人称视角,其黑箱性质和难以捉摸的自主学习特性无法促进形成由公正司法官和公开法律程序构成的权威审判景象,⑤加之它并不会与当事人面对面交流(即使这么做,当事人也并不期待一台机器会理解自己),因此它不能有效获取当事人的信任。

其次,预测性算法通过技术壁垒为诉讼公开性划界,进一步削弱了诉讼参与性。自然公正原则要求任何案件的裁判结果必须具备充分详实的推理论证,以揭示其背后的决策逻辑,而"算法黑箱"

① "Justice-1"不仅缜密推理了法条,而且开创性提出并计算了案件的"情感损失系数"和"商誉折损概率",其甚至在判决书中援引了一项连原告团队都未曾察觉的被加密的关键证据、分析出了该证据所涉人员的心理波动曲线并将其用以量化当事人的主观恶意程度。见《全球首例!AI法官主审民事案,司法界大地震》,今日头条,https://www.toutiao.com/article/7487040744625127987/,2025年10月1日访问。

② 劳佳琦:《AI助力,实现更有效率的司法正义》,《光明日报》2025年4月26日,第5版。

③ Tom R. Tyler & Hulda Thorisdottir, *A Psychological Perspective on Compensation for Harm: Examining the September 11th Victim Compensation Fund*, DePaul Law Review, Vol. 53, No. 2, 2013, pp. 355-392.

④ Frederick Wilmot-Smith, *Equal Justice*, Harvard University Press, 2019, p. 1.

⑤ Adrian Zuckerman, *Artificial Intelligence: Implications for the Legal Profession*, *Adversarial Process and Rule of Law*, *Adversarial Process and Rule of Law*, Law Quarterly Review, Vol. 136, 2020, pp. 427-453.

问题[①]为司法决策蒙上了面纱。预测性算法以输入数据、处理并计算、输出结果为主要步骤，但此类自动化决策是秘密进行的，算法对生成结果的解释也并不足以使当事人充分了解其依据与逻辑。即使算法决策侵犯了当事人合法权益，后者也无法知悉“算法心证”，法律甚至未提供渠道以便当事人提出质询或进行救济，这何尝不是一种“算法暴政”?[②] 如在我国2025年上海市青浦区张某盗窃案中，AI量刑辅助系统建议对被告判处2年实刑，其辩护人要求公开算法参数，被法院以“技术秘密”为由拒绝。[③]

鉴于算法结论往往对被告权益产生实质性影响，被告对裁判其责任的依据或证据进行检视和质疑的权利变得尤为重要。[④] 我国法律亦规定了司法机关的权利告知义务以及诸如阅卷权等保障被告知情权的法律制度。然而，现实中被告因无法触及算法的核心分析环节而处于信息劣势地位，难以对算法分析过程与结论准确性提出质疑，进而无法在知情情况下进行有效抗辩。在这个意义上，被告很大程度上被剥夺了陈述、辩解或补充的程序性权利。作为充分参与诉讼、有效实施抗辩的前置条件，被告方有权并需要掌握充足信息。但这种通过限制诉讼公开性进而淡化甚至剥夺参与性的自动化决策机制，使被告前述相关程序权利难以落到实处。

四、预测性算法的人权保障进路及其对我国的启示

杰克·巴尔金(Jack M. Balkin)指出，对算法规制的核心不在于算法技术本身，而在于算法使用者与被算法影响者，算法治理本质上是由人使用特定分析决策技术对人的治理。[⑤] 因此，在探求如何规制算法引起的失序时，应将思考重点放在因算法而发生权力配置偏差的社会关系上。技术的不稳定性最终会在执法终端显现，因此，法律必须修正执法终端的偏差。[⑥]

在预测性算法日益嵌入刑事司法的现实背景下，构建并完善具有中国特色的刑事诉讼算法规范体系，不仅有助于回应技术发展对人权保障提出的新挑战，更有利于顺应司法智能化改革的历史进程，防止技术逻辑凌驾于程序正义之上，从而为人工智能在司法领域的有序发展、审慎应用与可创新提供制度支撑。

(一)提高预测性算法准入门槛并建立算法解释机制

一是提高预测性算法准入门槛。在决定将预测性算法引入刑事司法之前，必须对其进行严谨调研与检验，否则会在整体层面增加预测性算法的失序概率，进而提高侵犯刑事被告程序权利的风险。为防止算法失序侵蚀程序正义，应在算法准入与运行层面建立严格、可执行的制度屏障。在准入阶段，应对算法的司法适用目的与潜在权利风险进行事前评估，防止未经验证的算法直接嵌入诉讼程序；在运行前审查阶段，通过司法系统内部监督与独立外部评估相结合的方式，重点检验算法

① 分为“有意的不透明”“无知的不透明”和“内生的不透明”风险，详见 Bruno Lepri et al., *Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Open Challenges*, Philosophy & Technology, Vol. 31, No. 4, 2018, pp. 611-627.

② Bruno Lepri et al., *The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good*, in Tania Cerquitelli, Daniele Quercia, Frank Pasquale eds., Transparent Data Mining for Big and Small Data, Springer Press, 2017, p. 5.

③ 《刑事司法AI化危机：算法量刑建议的公开抗辩与程序抵制》，北京浩伟律师事务所网，https://bjxsbhls.com/zixun/28401.html，2025年12月26日访问。

④ Michael Brenner et al., *Constitutional Dimensions of Predictive Algorithms in Criminal Justice*, Harvard Civil Rights-Civil Liberties Law Review, Vol. 55, No. 1, 2020, pp. 267-310.

⑤ [美]杰克·巴尔金：《算法社会中的三大法则》，刘颖、陈瑶瑶译，《法治现代化研究》2021年第2期，第188页。

⑥ 蒋勇、张晓华：《算法时代预测性警务的兴起及其风险规制》，《公安学研究》2023年第6期，第13页。

的合法性与对被告权利的影响;在事后阶段,建立常态化复评与退出机制,一旦发现算法偏离公正裁判目标或产生系统性不利后果,应及时限制或停止对算法的使用。唯有通过这种前置把关、动态监督与可退出的规则设计,方能在承认算法工具属性的同时,有效控制其对刑事司法权力结构与人权保障造成的风险。

二是以算法解释机制为核心回应"算法黑箱"风险。为应对"算法黑箱",必须坚持刑事诉讼保护人权的立场,实现算法分析运行的全过程留痕,确保其决策逻辑可被追溯。预测性算法既已进入刑事诉讼领域并服务于定罪量刑程序,那么其必须受到来自刑事诉讼宗旨的约束。为此,有必要通过制度设计确保算法决策具备最低限度的可解释性与可追溯性,使司法人员与被告能够理解算法结论的生成依据,而非被动接受其结果。同时,裁判者在具体使用预测性算法时应当避免输入"一步到位"式的笼统问题,而应当将所欲解决的最终问题拆分为若干个更为细致具体的输入指令,并结合案情不断补充递进提问时的细节条件。虽然这样并不能完全规避算法运行黑箱,但至少可将对最终问题的思考方向和探寻路径部分掌握在人类手中。无论如何,保障人权是刑事诉讼的核心价值,该价值不应因算法的私人属性或商业利益而让步。算法解释机制一定程度上可以缓解算法私人性和秘密性对被告权利的不利影响。

(二)设立司法人员的算法使用说明义务

所有纳入算法的数据都来源于历史,因而算法的思考本质同司法人员的经验法则并无二致。算法生成的评价基于聚类数据作出,往往因缺乏灵活性而形成"一刀切"的机械效果。算法产生的歧视与偏差亦需要得到司法人员的监督与纠正。

1.以审判为中心,建立"控方解释、辩方质询"机制

决策者过度依赖算法生成的概率性判断,将降低对被告个性化特征的重视程度,被告的发声亦会被削弱,该倾向不利于被告对决策的产生发挥作用。① 技术发展若仅赋能国家权力而忽视对个体权利的保障,被告方便会在刑事诉讼场域中处于技术和信息劣势。

就控方而言,当其使用预测性算法作出控诉策略,并试图据此影响关乎被告利益的司法裁量时,应相应强化其程序性告知与说明义务。可在制度上延展我国《刑事诉讼法》第120条第2款有关控方的权利告知义务规定,明确控方有义务告知辩方其针对预测性算法所享有的程序性权利,包括算法结论的检视权与质询权;与此同时,参照《中华人民共和国个人信息保护法》第24条第3款之规定,控方应告知并向辩方展示自己对预测性算法输入的指令,以及与算法互动的全过程,以避免算法结论成为不可质疑的"黑箱决策"。

就辩方而言,为进一步加强其诉讼主体地位,可设置审前算法使用异议制度。在证据开示阶段,若被告方有初步证据证明控诉机关所使用的算法存在错误,其可在审前向法院提出针对预测性算法的异议申请。法院对异议申请及其初步证据应该进行审查,确认算法存在瑕疵的,可依被告方之请求,要求控诉机关在本案中停止使用该算法。诉讼过程中,司法人员使用预测性算法且辩方认为确有必要时,应允许辩方委托专门技术人员对算法进行评估并针对算法结果发表意见,以辅助辩方有效质询与辩护。此外,在被告反对算法结论时,还应为其设置反馈与救济渠道。通过上述制度安排,促成算法介入背景下控辩结构的相对平衡。

① Emily Berman, *Individualized Suspicion in the Age of Big Data*, Iowa Law Review, Vol. 105, No. 2, 2020, pp. 463-506.

当下预测性算法已实现情报与信息的密集产出，算法在未来诉讼过程中将发挥越来越大的作用。鉴于控辩双方的辩论可能成为算法结论的外化，应预见并重视诉讼核心环节被算法逻辑支配的风险。因此，不论是提高控方的控诉成本，还是增加辩方的质询权利，在诉讼过程中的权力配置方面，审判职能应当对控诉职能进行有效制约和规范，同时切实保障辩护权的实质行使，确保审判在诉讼阶段处于核心地位。这亦顺应我国以审判为中心的司法体制改革趋势，符合我国《中共中央关于全面推进依法治国若干重大问题的决定》以及《关于全面推进以审判为中心的刑事诉讼制度改革的实施意见》等政策与司法文件的精神。

为实现庭审实质化，审判者需在认定证据、保护辩护权方面适应数字化趋势，发挥决定性作用。首先，全面严格审查刑事证据。目前，我国对证据认定仍采法定主义，然而诸如专家评估意见、大数据分析报告等大量由算法生成的信息数据正涌入法庭，成为一种认定案件事实的新兴因素。这些“机器证据”[①]形式众多、证据能力要素复杂，审判者往往难以对其进行归类及审查。证据的审查认定将成为影响刑事被告权益的关键，同时算法可能产生幻觉，因此审判者应更加审慎对待此类“机器证据”。在针对此类材料的审查规范尚未制定时，法官应将此类材料纳入质证体系，经由辩方质询，并聘请专家辅助判断，以严格标准对其进行认定。只有在“机器证据”的证据能力得到确证的情况下，[②]法官才可酌情将其纳入对事实的考量中。如此可以进一步明确控方的举证边界与证明责任，防止其借助预测性算法将风险评估结果转化为事实认定依据，从而降低算法使用对无罪推定原则产生冲击的可能性。

此外，法官应加强对认罪认罚案件的全面实质审查，确保被告认罪认罚意图真实且系基于平等协商作出，而非基于控诉机关借助预测性算法形成的不对等条件作出，并依法纠正明显不当的量刑建议。在辩护权保障方面，法官可酌情支持辩方的数据访问申请，允许其对预测性算法的数据来源、处理过程及分析方法进行质询；同时，应禁止控方以“数据倾倒”方式变相削弱辩护效果。必要时，法院还可要求预测性算法的开发人员到庭说明相关技术原理与决策逻辑。

2. 以法官释明义务确保审判独立

在司法决策中，防范自动化决策侵蚀司法人员的独立判断与客观性，需避免裁判结论过度依赖算法输出。核心在于明确算法决策在司法裁判中的功能定位，厘清其与司法决策之间的主从关系与边界。首先，在司法机关正式将预测性算法投入系统使用之前，应对算法覆盖的司法人员进行必要培训，培训固然需要指导司法人员正确且高效地使用算法，但更应当以技术祛魅、强化独立公正意识为重心。其次，应当在算法使用界面设置明显的风险警示，持续提醒司法人员算法可能混入错误数据与结构性偏见甚至产生“幻觉”，促使其批判性看待算法作出的预测评估结果，并有意识地将算法放置于决策从属地位。最后，在司法裁判阶段，必须警惕审判人员未经实质审查即直接援引算法分析结论，防止技术工具侵蚀司法裁量权的独立性。在法官决定将算法生成内容纳入判决书作为裁判说理依据时，应当强制其对算法生成内容的真实性、可靠性及逻辑合理性进行验证，并将验证结论与采纳的理由完整写入判决书。

结语

数字时代，预测性算法逐步嵌入刑事诉讼运行结构中，技术革新暴露的种种人权风险亟待正

① 该概念由安德烈亚·罗斯提出，详见 Andrea Roth, *Machine Testimony*, Yale Law Journal, Vol. 126, No. 7, 2017, p. 2000.

② 尤其在该证据由控方提出时，其不仅应符合“证据三性”标准，控方还应就该证据的真实性进行重点检验与说明。

视。无论从国际人权法还是刑事诉讼法理的角度看,预测性算法均难以为减损被告权益提供充分的正当性依据。在程序正义维度,算法通过先入为主的风险预判影响法官裁判,其放大"数字鸿沟"并使控辩两造格局的力量失衡,同时,算法与无罪推定原则精神相抵触,并持续弱化被告的诉讼参与权。

这些风险清晰表明,算法进入刑事程序的事宜必须被严格规制,算法永远不能替代人类对正义的判断与权衡。预测性算法在我国刑事司法体系应用前景广阔,前瞻性布局规制的制度意义重大。这不仅是应对未来技术挑战的准备,更是维护司法公正、坚守人权保障底线的必然要求。正如於兴中教授所言,"人工智能的限度应当是人的限度",①司法的核心始终是"人"的权利与正义。唯有以人权为标尺划定技术边界,才能让算法真正服务于司法公正,推动数字司法在高效与温度之间找到平衡,实现技术与法治的良性融合。

Application of Predictive Algorithms in Criminal Proceedings: Rights-Related Risks and Their Prevention

SUN Fabai, LIU Qinghua

(*College of Humanities and Law, Shandong University of Science and Technology, Qingdao* 266590, *China*)

Abstract: Against the backdrop of the digital era, predictive algorithms have been increasingly applied in criminal proceedings, giving rise to rights-related risks that urgently call for regulatory responses. The improper use of predictive algorithms entails tangible infringements on human rights, and does not reconcile with the fundamental principles of international human rights law, as embodied in the *Universal Declaration of Human Rights* and the *International Covenant on Civil and Political Rights*. Specifically, predictive algorithms may lead judges to render unjust decisions based on preconceived risk assessments, and exacerbate imbalances between the prosecution and the defense due to asymmetries in algorithmic resources and capabilities, thereby undermining the effective protection of defendants' rights. Besides, it may generate inherent tensions with the presumption of innocence, while automated decision-making may further weaken defendants' rights to information and meaningful participation in proceedings. To mitigate these risks, it is necessary to regulate the use of predictive algorithms in a prudent and principled manner by upholding the baseline of judicial fairness and rights protection. This requires imposing strict admission criteria for predictive algorithms, establishing mechanisms for algorithmic explainability and traceability of reasoning paths, and constructing a trial-centered framework for prosecutorial-defense balance, so as to consistently confine predictive algorithms to an auxiliary and subordinate role.

Key words: criminal procedure; predictive algorithms; procedural justice; risk prevention

(责任编辑:董兴佩)

① 於兴中:《算法社会与人的秉性》,《中国法律评论》2018 年第 2 期,第 59 页。