

智慧司法场域下人工智能辅助精准量刑的理论供给与规范应用

——以法官行使刑事自由裁量权为切入点

陈梓畴

广东理工职业学院, 社会工作学院, 广东 广州

收稿日期: 2025年9月9日; 录用日期: 2025年9月18日; 发布日期: 2025年10月13日

摘要

面对“案多人少”与“类案不同判”的双重压力, 经验式量刑已难以满足公正与效率并重的司法需求。研究以规范分析与技术法学交叉的方法, 聚焦人工智能介入量刑之理论与实践, 提出以“数据-规则-算法”三维映射重塑法院量刑结构: 在责任刑-预防刑区分框架下, 机器完成量化计算, 人类保留价值判断, 实现“量化可以量化的、裁量必须裁量的”权责配置。进而研究揭示数据偏差、算法黑箱、结果依赖与信息安全四重风险对实体公正和程序正当性的系统性冲击, 对此针对性地提出四项规范化路径。最终研究指出, 人工智能应定位为裁量支持而非裁量主体, 唯有在技术理性与司法理性深度耦合、实体与程序并重的前提下方能真正实现精准量刑与“同案同判”, 为我国智慧司法提供可复制、可推广的制度蓝本。

关键词

智慧司法, 人工智能, 精准量刑, 刑事自由裁量权, 可解释性

Theoretical Supply and Normative Application of AI-Assisted Precision Sentencing from the Perspective of Smart Justice

—Focusing on the Exercise of Criminal Discretion by Judges

Zichou Chen

School of Social Work, Guangdong Polytechnic Institute, Guangzhou Guangdong

Abstract

Facing the dual pressures of heavy caseloads with insufficient judicial resources and “different judgments for similar cases”, experience-based sentencing has struggled to meet the judicial needs of balancing fairness and efficiency. This study employs a cross-methodology of normative analysis and technical jurisprudence, focusing on the theory and practice of artificial intelligence (AI) intervention in sentencing. It proposes a three-dimensional mapping of “data-rules-algorithms” to reshape the court’s discretionary structure: under the framework distinguishing between liability punishment and preventive punishment, machines complete quantitative calculations, while humans retain value judgments, realizing the power-responsibility allocation of “quantifying what can be quantified and exercising discretion where necessary”. The study further reveals that four risks—data bias, algorithmic black box, result dependency, and information security—systematically impact substantive justice and procedural legitimacy. In response, four standardized pathways are proposed. The study finally emphasizes that AI should be positioned as discretionary support rather than the discretionary subject. Only under the premise of deep integration between technical rationality and judicial rationality, and equal emphasis on substance and procedure, can precision sentencing and “same cases, same judgments” be truly achieved, providing a reproducible and promotable institutional blueprint for China’s smart justice.

Keywords

Smart Justice, Artificial Intelligence, Precision Sentencing, Criminal Discretion, Explainability

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 问题的缘起

作为我国刑事司法现代化的核心，量刑规范化改革自 21 世纪初启动，旨在回应全球化浪潮并根治“类案不同判”的痼疾，以在日益增长的案件量与有限司法资源之间寻求公正与效率的平衡。尤其在 2014 年最高人民法院《关于常见犯罪的量刑指导意见》全面施行，及随后认罪认罚从宽制度大规模推行的背景下，量刑建议的精准性与统一性被提至前所未有的高度([1]: p. 148)。然而，这一改革路径在本质上仍属于一种经验式量刑的范畴，其内在的有限理性与主观能动性，使得法院刑事自由裁量权的行使易在“恣意性”与“机械性”两个极端之间摇摆，难以完全克服量刑失衡的风险，从而未能彻底回应社会公众对司法公正的深切期待。

在此背景下，以人工智能为核心驱动的智慧司法建设应运而生，并深刻催生了智能式量刑这一创新实践([2]: p. 98)。从历史维度审视，国际上不乏早期计算机辅助量刑的尝试，虽囿于技术或理论的局限性而未能大规模普及，但其对量刑精准化的不懈追求却与当代人工智能辅助量刑理念一脉相承([3]: p. 61)。如今，依托于大数据、云计算及先进机器学习算法(如 DeepSeek-R1 等大模型)的强大支撑，智能量刑辅助工具能够实现案件的自动化分析、类案精准推送与量刑偏离预警，使量刑实践从个体经验判断迈向更为科学的智能辅助预测时代，无疑为法院优化刑事自由裁量权的行使提供新的技术可能([4]: pp. 192-193)。

然而，技术的跃迁并无法完全消弭其悖论。首当其冲的是历史回顾性数据本身可能内含直接或间接

地歧视性信息,人工智能在学习这些数据时可能重现甚至加剧既有的偏见,导致“假阴性”或“假阳性”的预测结果,严重侵蚀量刑实体公正的基础([5]: p. 143)。当算法运行机制不公开、决策逻辑不可解释时,被追诉人对程序参与、有效辩护的权利可能被架空,法院裁判的说理性与公信力亦将受到质疑。更有甚者,对智能量刑结果的过度依赖,导致法官刑事自由裁量权的异化与削弱,使其主体地位逐渐让位于冰冷的算法,形成一种“数据主义司法观”,进而模糊刑罚目的位序的考量,使量刑实践陷入技术规训的囹圄。

鉴于上述困境,当前学界对人工智能辅助量刑的研究,多有探讨其应用场景、风险规避及程序机制构建[6],但往往囿于宏观抽象或片面关注,未能对人工智能如何精准赋能、规范法院刑事自由裁量权缺乏系统而具象的理论供给。

简单的程序性规制,难以从根本上提升智能量刑结论的合理性与科学性,亦无法阐明其精准预测功能是否真正契合量刑规范化的初衷,能否经受住刑罚理论的严苛检验([1]: p. 149)。

2. 人工智能辅助精准量刑的逻辑证成

智慧司法场域下,人工智能技术辅助精准量刑并非旨在取代法院在刑事案件中的自由裁量权,而是通过深度学习与数据分析,弥合传统经验式量刑固有的认知偏差与局限,从而构建更为科学、透明、公正的量刑决策机制。

(一) 经验裁量失衡与法院刑事自由裁量权量化的内在需求

我国量刑规范化改革已历经十余载,并伴随《关于常见犯罪的量刑指导意见》的全面施行及认罪认罚从宽制度的深入推行,在统一法律适用、规范裁量权方面取得显著成就([1]: pp. 148-149)。传统“估堆式”量刑方法高度依赖法官个体的业务素质与办案经验[7],使得量刑结果呈现不确定性与非一致性,加之“案多人少”的现实压力,法官有限的精力与时间难以支撑对海量案件的精细化处理,无形中加剧“类案不同判”的风险([3]: p. 63)。

不可否认的是,经验式裁量模式的失衡折射出人类认知的有限理性。法官虽具有丰富的法律知识与实践智慧,但其对整个司法系统量刑范式的全面把握以及从历史错判中汲取经验的能力,仍受制于个体经验与记忆的局限([1]: pp. 149-151)。譬如,在许霆案¹一审判决所引发的社会广泛争议,恰恰暴露出仅满足形式合法性,而缺乏实质合理性、脱离社会情理的判决,即便符合罪刑法定,也可能与罪责刑相适应原则相悖([8]: pp. 138-140)。

因此,将刑事自由裁量权中可量化、可规则化的部分加以数字化、算法化,将那些隐性的、个体化的经验转化为显性的、共享的司法理性,为法官的裁量提供更为坚实、统一的基准,以弥补人类法官在信息处理广度、深度与一致性上的不足,已成为刑事司法现代化进程中一项内在且迫切的需求。

(二) “数据-规则-算法”作为三维分析框架的理论价值与实践功能

在智慧司法场域下,人工智能辅助精准量刑的应用逻辑在于将法院传统经验式裁量结构,重新映射为一套以“数据-规则-算法”三维分析框架的协同运作体系。该三维框架不宜简单视作一个新模型,而应被定位为理解与构建人工智能辅助精准量刑系统的核心分析工具,其本质上是法官内隐的裁量经验通过技术手段外化为可操作的规范,从而提升司法裁判的统一性与可预测性。

司法大数据,尤其针对已生效刑事裁判文书,蕴含着海量有价值的量刑规律与司法共识。人工智能

¹一审(2007年11月,广州市中级人民法院)罪名:盗窃罪(根据《刑法》第264条,盗窃公私财物,数额特别巨大)。认定:许霆明知ATM机系统故障,超额取款且占为己有,其行为构成盗窃罪。量刑:无期徒刑,剥夺政治权利终身,并没收个人全部财产。

二审(2008年1月,广东省高级人民法院)裁定:一审判决认定许霆犯盗窃罪事实不清,证据不足,撤销一审判决,发回重审。

一审重审(2008年3月,广州市中级人民法院)改判:许霆犯盗窃罪,判处有期徒刑五年,并处罚金2万元。法院理由:一审定罪准确,但原判量刑显著超过行为的主观恶性和社会危害性,不符合罪责刑相适应原则,故予以改判。

通过自然语义识别等先进手段对非结构化文本数据进行深度挖掘与要素提取,将其转化为结构化、可计算的量刑要素(如犯罪事实、罪名、量刑情节、宣告刑等),进而构建起庞大的司法知识图谱[9]。简言之,这种全量分析能力是对法官集体智慧的客观记录,是超越个体法官有限认知的“司法大智慧”[10]。

进而,人工智能通过归纳学习识别量刑要素间的复杂关联性,将法官对量刑情节的取舍、权重的考量等隐性经验,转化为更为具体的、可量化的规则[11]。这实质上是对量刑指南中所蕴含的演绎算法的具象化与精细化,使得量刑过程从模糊的“估堆式”转向精细的“阶段化”与“类型化”处理实现了量刑规则的再编码,例如对于常见罪名或认罪认罚案件给出更为明确的量刑起点、基准刑与情节加减量范围([12]: pp. 89-90)。

在归纳学习的基础上,人工智能算法决策凭借“量化规范逻辑”和“经验规范逻辑”,实现案件信息向量化、属性表示拼接以及刑期类别概率分布的映射,完成快速精密的计算([13]: p. 144),进而通过流程化裁判方式,将定罪、量刑要素加以分析、计算,裁判结果与已决案件进行比对,使得法院的裁量结构从传统的人脑驱动,转向数据驱动、规则指引、算法赋能的现代化模式,为精准量刑提供坚实的逻辑基础。因此“数据-规则-算法”被定位为核心分析工具,其价值不仅在于理论层面的范式革新,更在于司法实践场域的结构重塑。

(三) “同案同判”价值导向下的人机协同权力再平衡

“同案同判”作为司法公正的朴素追求,在人工智能辅助精准量刑的背景下其内涵正被重新建构。当然,必须肯定的是绝对意义上的“同案”在现实中并不存在。其所强调的是司法实践中对于相似案件,应在罪责刑相适应原则下保持量刑方法的统一性与结果的合理一致性,以避免因法官个体差异而导致的司法不公([14]: p. 194)。

人工智能辅助量刑在此价值导向下,通过对海量司法数据的深度分析识别并提取出案件的核心量刑要素,并根据已生效判决的量刑规律,为待决案件提供精准的量刑预测意见。这使得法官在行使自由裁量权时获得来自“集体司法理性”的强大支撑[15]。其实际上是一种经过大数据验证的“经验性应得”的体现,反映了社群对特定犯罪行为的普遍正义共识,从而在基数比例(即特定罪行的刑量是否合比例)层面为法官提供了更为精确的量刑基准([1]: pp. 153-155)。

这种协同并非人工智能对法官权力的替代或侵蚀,而是一种权力再平衡。²即人工智能系统作为“辅助工具”完成量刑中“刑之量化”的部分,即通过算法对案件事实与罪名、刑事责任之间的量化关系进行精确计算,提供个罪的量刑基准。使得法官能够从繁重、重复性的量化工作中解脱出来,将有限的精力与智慧集中于“刑之裁量”的核心环节,对个案特殊性、复杂酌定情节以及人身危险性等难以完全量化因素的价值判断([4]: pp. 195-197)。

法院刑事自由裁量权得以在两个维度上实现优化:一是理性限制,人工智能的介入能规避法官在裁量过程中可能受到的非理性因素(如偏见、锚定效应等)干扰,确保判决结果的稳定性与公正性([12]: p. 90);二是精准赋能,系统提供量刑预测意见,使得法官的裁量不再是无根基、无标准之裁量,而是在明确的量化基准上进行的科学裁量。最终目标在于克服“同案异判”的弊端,使每一份判决都能在合理限度内最大程度地实现社会正义的价值追求。

(四) 责任刑-预防刑双轨视角下精准量刑的合法性阐释及人工智能的审慎应用

人工智能辅助精准量刑的合法性须回归到刑罚目的论的核心问题——即在报应刑(责任刑)与预防刑的双轨价值体系下,如何确保人工智能的介入能够更好地实现刑罚的正义、公平与人道[16]。然而,责任刑与预防刑的理论二分在司法实践中常呈现边界模糊性,同一案件情节往往同时关涉已然之罪与未然之

²《最高人民法院关于规范和加强人工智能司法应用的意见》(法发〔2022〕33号,2022年12月9日发布)明确了“辅助审判原则”,强调人工智能不应削弱用户自主决策权,这正是对“人机协同”中“人主机辅”责任分配格局的官方确立。

险，这为人工智能的技术逻辑构成严峻挑战。其依赖历史数据与量化模型的固有特性，难以处理此类复杂的规范评价交叉性问题，易将归责问题简化为统计运算，甚至固化既有的偏见。

鉴于该复杂性，人工智能的司法应用须极为审慎，并依据双轨功能差异确立清晰的适用边界。在报应刑的框架下，报应刑强调刑罚的严厉性应与犯罪的严重性相适应，追求“罪有应得”的积极比例原则([1]: p. 154)。人工智能借助海量判决数据，能够精确识别和量化反映犯罪严重性的客观要素(如犯罪数额、犯罪手段、犯罪结果等)，从而为责任刑提供精准的基准值。该类“向后看”的算法在收集和处理责任刑数据方面具有天然优势，因其涉主观因素相对较少，推理过程更易被机器学习。使报应刑所追求的“恶有恶报”的正义在量化与一致性上得到了前所未有的提升，通过数据驱动的经验性应得，将社群对犯罪谴责的共识具象化。

而在预防刑³的框架下，人工智能则需更为审慎处理人身危险性的判断，并在刑罚目的位阶上遵循“附条件优先”原则。然而，对人身危险性的评估涉及犯罪人的主观心态、社会背景、改造可能性等复杂因素，这难以完全量化且蕴含历史数据中的歧视偏见。因此，在现阶段，人工智能在预测预防刑方面的能力仍受到限制，人身危险性的判断应更多地保留在经验式量刑的范畴，由法官凭借其独特的价值判断能力与人文关怀来综合考量。

为实现人工智能的审慎应用，应建立“以人为本、技术为辅”的功能互补范式，并辅之以相应的技术保障机制：在理念上，明确人工智能仅作为辅助工具，绝非对法官裁量权的替代或侵蚀；在技术上，可以探索引入“提示机制”，要求算法标注所依据的关键情节是作用于责任刑或预防刑，通过该种提示增强法官对算法决策的审查能力，确保其技术理性最终服务于司法理性([2]: pp. 109-116)。

人工智能辅助精准量刑的合法性阐释，主要聚焦于对已然之罪的客观归责(责任刑)，法官则将其核心裁量权集中于对未然之罪的价值判断与个体化考量(预防刑)，从而形成一种“以报应优先的预防区分”的刑罚目的位序([1]: pp. 156-157)。而面对责任刑与预防刑在实践中的边界模糊性，需通过确立“功能互补”的人机协同范式，并借助“提示机制”等相应技术机制增强法官对算法过程的审查([2]: pp. 109-112)。因此，在责任刑-预防刑框架下，精准量刑的合法性及人工智能的审慎应用确保精准量刑追求“结果”的精确统一，更兼顾“方法”的科学合法，实现技术理性与司法理性的深度融合与相互赋能。

3. 人工智能辅助精准量刑的运行难题

(一) 数据偏差对自由裁量实体正当性的侵蚀

人工智能介入量刑环节的根本动力在于用数据再现过往司法经验、用模型预测未来裁判走向，但正是这一“数据-模型-预测”链条的首端已暗藏裂隙。

我国现阶段的司法数据呈现出典型的“块状缺失”与“结构性失真”。一方面，判决书公开虽突破以往信息壁垒，但涉未成年人、涉国家安全、涉商业秘密等豁免条款，致使大量案件缺席样本池；另一方面，即便已有文书被采集，关键字段的抽取也高度依赖人工标注或基于关键词的简单抓取，诸如被告人悔罪态度、受害人过错程度、地方社会风俗等对量刑关键性情节以非结构化文本或维度缺失的方式被过滤掉。迫使模型训练过程中借用可量化但相关性并不充分的替代变量(proxy variable)进行推断，而替代变量往往与特定地区、职业、族群的社会经济标签高度重合，由此形成“技术性歧视”的暗流。而历史数据本身已深受地区经济发展水平、社会治安形势以及阶段性刑事政策(如严打、宽严相济)摆动影响的前提下，算法若再机械地赋予同等权重，便会将过去的结构性不公“冷冻”在模型参数中并反向输出到未来的裁判实践。其直接后果是：一旦法官对模型评分产生路径依赖，便易出现“高风险评分 = 重刑”“低

³预防刑旨在通过威慑、矫正、剥夺犯罪能力等手段，防止犯罪人再犯或一般人犯罪，追求刑罚的功利性目的(减少犯罪与维护社会秩序)。

风险评分 = 轻刑”的自动贴标效应，而忽略重新审视该评分在个案中的适足性。

当算法将“同案不同判”简化为形式上的“同案同分”，自由裁量对个别正义与罪责刑相适应原则的动态把握便被“分值 - 量刑区间”映射所取代。更为隐蔽的是，技术团队在模型开发阶段的特征筛选、样本再平衡、阈值设定均内嵌在代码与服务器之中，其价值取向并不在司法体系的显性监管范围内却能实质左右量刑建议的走向。于是，看似“去经验化、去情感化”的算法，实则把编程者的世界观与历史数据的偏见层层叠加后投射给法官，而后者若缺乏足够的技术辨识力与再验证机制，就会在无意中把量刑实体正当性的控制权拱手让渡。

(二) 算法黑箱造成的程序参与与可解释性缺位

若是说数据偏差蚕食量刑正义的“内容”，那算法黑箱则动摇了程序正当性的“形式”支柱。深度学习模型能够输出近似人工经验判断的高维度结果，恰源于其极端复杂的参数网络——数以百万计的权重在隐藏层中交互迭代，任何单一节点难以被线性还原——天然构筑起一个难以被外部窥视的“技术封闭领域”([17]: pp. 113-114)。

对于控辩双方而言，量刑环节的核心诉讼权利是“知情 - 质证 - 辩驳 - 说服”，而当算法充当量刑意见提供者时，第一步“知情”便已陷于困境。算法开发者以商业秘密、网络安全或防止模型被“对抗攻击”为由，拒绝公开源代码与完整训练数据；即便披露关键权重或部分样本，对非技术主体而言仍属“数字噪声”。这种情形下，当事人仅看到一份简化版的“风险报告”或“量刑区间建议”，却无法追问“何以得出”、“为何权重如此”、“若改动某一变量，结论会否改变”等技术细节，司法程序性对抗机制骤然失灵。更进一步，法官自身作为裁判主体缺少对神经网络原理的深度理解，只能借助外部专家或开发者的说明文件“转述”算法理由，从而形成一种“说理替代”——判决书中出现日益频繁的经系统评估，被告人再犯风险较高、智能平台建议量刑 X 年等表述，其背后并无可供审查的逻辑链条[18]。

结果便是一方面，传统意义上辩护、法庭辩论、证据质证在面对高度技术化的量刑意见时被边缘化；另一方面，上诉、再审等外部救济程序也因缺乏可复盘的推理路径而难以评估一审采信算法是否存在事实认定或法律适用错误。司法公信力依赖的是“易懂可辩、能检能复”的公开过程，而算法不可解释性恰好堵塞了此条路径。久而久之，当事人对司法正义的感知将从“被倾听、被理解”滑向“被告知、被接受”，裁判从“理由治理”退化为“技术治理”([3]: p. 64)。因此，若无配套的解义务、接口可视化与独立第三方审计制度支撑，算法黑箱剥夺了诉讼主体的程序性“在场”，还可能在外观上制造出一种“科技保障公正”的幻象，却在制度深处掏空自由裁量赖以安身立命的程序理性。

(三) 结果依赖引发的裁量权异化与司法失衡

在人机协同框架下，人工智能量刑系统常被宣称仅发挥“参考”“校正”作用，但技术便利性与组织绩效考核双重压力使法官易陷入“算法锚定”与“路径锁定”的双重依赖([2]: pp. 107-108)。一旦系统能在几秒钟内给出精确到月份、甚至天数的量刑区间，而人工书面检索需耗时数小时乃至数日，法院内部的案件流转机制便倾向将算法输出认定为高效与规范的象征；同时，审判管理对“类案同判率”、“发回改判率”设定的量化指标，客观上也鼓励承办法官优先采纳算法建议以降低事后纠错风险。

如此一来，原以对事实与法律的主观认定为核心的裁判活动，被技术评分与量化分档重塑为“输入 - 输出 - 审核”的流水线，法官的自由裁量不再围绕刑事政策与罪责刑相适应原则展开，反而围绕系统给出的建议与最终判决差距这一绩效指标进行自我校准。长此以往，个案适用中的主观能动性被挤压，审判职能蜕变为对技术建议的形式化背书；法官角色由“法律解释者”异化为“技术执行者”[19]。

更何况，算法由第三方公司开发、由政府采购部门引入、由信息化中心维护，刑事量刑这一典型的公权力核心环节便在事实上对多重外部主体敞开。由于这些主体并不受《法官法》中独立行使审判权之义务约束，且其决策逻辑隐藏在合同条款、源代码与服务器权限背后，司法权的行使链条呈现技术外包

化和责任碎片化趋势。当量刑结果出现瑕疵，追责路径容易陷入“系统建议－法官采信－开发公司算法－数据提供者”多节点循环归责的迷宫，易诱发各主体互相推诿、难以形成有效纠偏。

若上下级法院部署同一算法平台，技术同源将导致上诉机关与原审机关在量刑基准上一致化，实质上削弱了两审制的纠错能力([13]: pp. 149-150)；若控方亦使用同一系统生成量刑建议并在庭审中提交，控审技术同轨又可能放大“控审合意”偏向，挤压辩护方的发言空间，形成事实上的程序失衡([17]: p. 113)。由此可见，结果依赖不仅稀释法官自由裁量的价值判断功能，更在制度层面引入对审级独立、职权划分乃至诉讼对等原则的系统性冲击。

(四) 信息安全隐患对司法公信力的冲击

与传统纸面卷宗不同，人工智能量刑系统在“端－管－云”全链条运行中沉淀了海量高敏感度数据：既包括涉案人的身份信息、既往犯罪记录、精神状态评估，也包括证人联系方式、司法工作人员账号权限等元数据。

按照我国《网络安全法》《个人信息保护法》以及等保 2.0 的分类分级要求，数据大多处于“敏感个人信息”与“重要数据”双重保护范畴，但实际操作中仍面临三类突出风险：

其一，存储侧的集中化导致“一点突破、全域失守”。目前多地法院将省级或地市级司法大数据汇聚于同一云端节点，一旦遭遇勒索软件攻击或内部人员权限滥用，数十万份判决书及原始证据链可在数小时内被批量窃取、外泄或篡改，直接威胁案件当事人及证人安全。

其二，训练侧的“数据污染”与“对抗样本”风险。黑客或利益相关者可通过向开放接口注入伪造样本、扰动权重或添加对抗性噪声，悄无声息地改变模型对特定群体或个案的评分阈值，从而对量刑结果产生系统偏移。

其三，应用侧的“交叉识别”与“算法复用”风险。跨部门数据共享虽提升了政务协同效率，却使当事人的犯罪记录、健康状况乃至财产信息可能在不知情情况下被商业机构或其他政府部门调用，进而引发二次歧视或差别化服务。上述安全隐患与司法公信力之间的关联具有传导放大效应：任何一次集中泄密、算法腐蚀事件都会迅速被社交媒体放大为“司法系统失守”的符号化新闻，公众对司法裁判独立性与可靠性的信任基础被迅速蚕食([1]: pp. 152-153)。

结果是，即便法院在单案中作出了合乎法律与事实的判决，信息安全阴影仍会使当事人和社会公众对幕后算法是否公正、数据是否被篡改等问题心存疑虑，进而将对司法系统的整体正当性产生外溢负面认知。

4. 人工智能辅助精准量刑的规范应用路径

(一) 数据质控循环与偏差预警：保障实体公正

人工智能辅助精准量刑的实体公正性主要取决于其训练与运行所依赖的司法大数据质量。鉴于数据偏差对自由裁量实体正当性的侵蚀，规范应用亟需从数据生成、汇聚到应用的全生命周期构建严格的数据质控循环与主动的偏差预警机制，以确保输入数据的“纯净”与“可用”。

其一，应确立并优化司法大数据样本的“理性经验”供给机制。正如研究指出的，理性的法官群体经验是提高数据质量的重要保障([5]: p. 145)。这意味着，人工智能量刑系统所学习“司法大数据”应着力提炼和吸收那些符合司法公正理念、体现罪责刑相适应原则的裁判经验，即在数据收集层面加强对历史错判、量刑畸轻畸重、或明显带有偏见的案例进行识别、标注与净化，避免“垃圾进，垃圾出”的困境，从源头上阻断歧视性决策的复制与放大。同时，应推动法官量刑理由输出的精细化与标准化，尤其是对酌定情节、个案特殊性的说理，将其作为高质量训练数据的重要来源，从而在数据层面有效校正可能存在的直接或间接歧视。

其二，仅仅依靠前端的“清洗”是不足够的，量刑系统还需具备对潜在偏差的主动感知与干预能力。包括对训练数据中可能存在的“假阴性”和“假阳性”预测进行风险评估与提示，例如，针对量刑情节赋值标准的不明确性通过“数据－规则－算法”的反馈循环，不断迭代优化量刑要素的量化标准和权重，使算法模型在处理责任刑与预防刑时，更精准反映刑罚目的位序的考量([1]: p. 152)。

同时，鉴于刑事制裁体系特点对量刑制度设计的实质性影响，量刑系统在偏离预警时应兼顾累犯、惯犯等情节的特殊性，逐步实现责任刑对最终刑量的“相对限制”向“绝对限制”过渡，以推动量刑实体公正的持续深化。

(二) 可解释算法与诉讼化听证：程序正当化机制重塑

为回应人工智能辅助精准量刑所造成的“算法黑箱”问题，及由此导致的程序参与和可解释性缺位，规范应用必须致力于实现量刑程序的正当化重塑，然亦需考量其操作成本与实施局限性。

一方面，应着力研发和应用具有“可解释性”的量刑算法与决策可视化工具，对此，需借鉴模型可解释性工具(LIME、SHAP 等)前沿技术，揭示模型在特定决策中的关键因素和逻辑推导，使非技术人员也能理解其决策依据。同时，应推广决策可视化技术，将复杂的算法决策过程以流程图、决策树、热力图等直观形式展示给用户。例如，在量刑预测系统中清晰呈现模型如何从法律条文、案例中筛选依据，并依据这些数据得出裁决建议，从而增强用户对模型决策的信任([5]: p. 151)。这要求技术开发者在系统设计之初嵌入“透明性因子”，使算法结果的生成具有可解释性、可回溯性，这与“规则科学”的理念一脉相承，即量刑规则被编码入系统时其内在逻辑和生成过程应能够被合理阐释。但须注意，“可解释性”存在层级之分，并非所有算法或决策场景均需要达到完全透明，对无需完全透明层级则可选择使用解释工具提供局部可解释性，而非一刀切地要求模型本身完全透明。

另一方面，应健全面向量刑算法的“诉讼化听证”程序机制，确保当事人的实质性程序参与与异议权。即赋予当事人以量刑算法解释权、发表意见权、质证权为核心的实质性参与权利。这意味着当事人有权知晓量刑算法的运作机理，并对算法决策中的不同看法发表意见，或基于技术水平等因素对量刑算法的可靠性进行质证。为此，可探索建立量刑算法决策系统技术人员出庭说明情况机制，并明确其责任承担，以保障算法运行及其决策的可理解性与可解释性([2]: p. 113)。同时，赋予当事人算法解释权等实质性参与权利时，也应建立分层级解释义务机制：对于常规性量刑建议，可提供标准化解释报告；对于重大疑难或争议案件时，则需要启动更深入的说明与质证程序，甚至要求技术人员出庭作证。此举既可保障当事人权利，也可控制司法成本。

(三) 权责分明与人机协同：裁量权动态校正模式

人工智能进入量刑领域的正当性始终建立在“技术工具论”而非“权力主体论”之上([2]: p. 103)，因此须以权责分明为前提，通过人机协同实现裁量权的动态校正。

技术层面，量刑链条可划分为数据抽取－指标建模－风险评分－结果呈现四个节点，其中前两环节侧重对客观事实与法条要件的量化解析，后两环节则不可避免地嵌入价值判断与社会评价[20]。

为防范工具理性对价值理性的吞噬，可将量刑决策再阶段化为机器量化区、人机共治区和人类裁量区三阶段：

- ① 机器量化区专注量刑起点、法定加重时长等事实维度的规范化；
- ② 人机共治区针对量刑情节作初步评分，但系统仅提供风险阈值和区间提示；
- ③ 人类裁量区留给法官对责任刑、预防刑及个案特殊性进行终局评估并给出最终刑期[21]。

如此设置可在制度架构上确立“量化可以量化的、裁量必须裁量的”分水岭。配套措施方面，应在法院信息化平台内嵌入算法建议－法官输入差异量化监测模块，对同一法官在一定周期内对算法建议的平均修正幅度、修正方向进行回溯性统计。当差异值长期趋零或呈单向同化时，系统自动触发警示并要

求承办法官撰写原因说明,既避免懒惰依赖,又为研修与能力培训提供数据反馈。

考虑到裁量的最终责任落在法官身上,判决书必须引入“算法引用”专栏,区分系统建议、法官考量与最终采纳三类内容并附逻辑阐释,确保裁判说理的主体仍是人而非代码。与此同时,建议将智能量刑报告依《刑事诉讼法》类比认定为具专门知识的人所出具意见,进入证据体系后须经控辩双方交叉质询,法院方可对其证明力进行确认。此举一方面将技术成果置于程序性阳光下,另一方面促使系统研发商积极强化可解释性与可质证性,从根本上抑制算法即真理的思维惯性。

(四) 分级授权与安全立法: 信息保护与公信力提升

司法数据与算法模型一旦成为数字时代的“生产要素”,便必须置于严格的分级授权与安全立法框架下,否则任何一次泄露或操纵都足以撕裂公权力的诚信壁垒。

对此,应当对司法数据实施“纵向分级、横向分域”的双维授权模型:

纵向分级依据数据敏感度、关联度与可逆匿名化程度,将原始卷宗、调查报告、未成年人信息等划为高度敏感层,仅限案件承办法官与经特别审查后授权的系统管理员访问^[22];

横向分域则按照业务流程将数据拆分为审判域、检察域、侦查域、社会化服务域四个封闭子网,通过零信任架构与多方安全计算完成必要交互,既避免“大集中”产生的单点脆弱性,也阻断跨域滥用链条。

在法律层面应以专门立法形式确立“算法可审计义务”和“国家安全豁免例外”,前者要求所有司法类算法在上线前接受独立第三方的源代码托管与可解释性评估,并每年度提交差异化影响报告,供全国人大及其常委会指定的数字法治监督机构审查;后者则针对重大公共安全、国家安全场景设定例外通道,确保在极端情形下司法机关能够及时冻结、备份或强制接管算法系统,阻断外部恶意入侵。

对算法偏见与安全漏洞的责任承担应引入“举证责任倒置 + 损害惩罚性赔偿”组合机制。当受害人能够证明算法存在显失合理的差别化对待时,系统开发商与部署机构需自证数据与模型无缺陷;若最终确认存在过错,则按照损害后果进行1~3倍惩罚性赔偿,并处以相应的行政罚款或刑事责任,以经济激励倒逼技术合规。为确保司法公信力,人民法院应建立判决数据脱敏复用沙盒,允许学术机构、媒体和社会组织在严格审查下对历史判决与算法表现进行再验证,及时公开安全事件与纠正措施。

5. 结语

刑事诉讼作为“智慧司法”转型初期的主阵地。在以“智能量刑算法”为典型的智慧司法改革面前,传统的法律概念(如“法官”“事实”“经验”甚至是“同案同判”)被赋予新的含义,由此开启了司法裁判的“技治主义”时代^[11]: p. 199)。基于智能量刑辅助技术在我国刑事司法实践中广泛应用和量刑程序改革的推进,相关程序研究需从宏观层面深入至人工智能辅助量刑这一具体场域^[2]: p. 116)。在此前提下,针对认罪认罚案件的人工智能辅助预测量刑,更具有迫切的现实需要。但须申明的是,不管在传统司法模式还是当下智能司法时代中,人类在量刑中的主体地位没有改变,只是在实现路径和方法上有所不同^[12]: p. 95)。对此,研究并不满足于“辅助”这一笼统定位,也不将人工智能仅视为提高效率的技术工具,而以法院刑事自由裁量权为切入,深入剖析智慧司法场域下人工智能辅助精准量刑的逻辑证成与规范应用之道,致力于从实体与程序的复合维度,探求在算法时代如何重塑量刑理念、优化裁量方法、健全制度保障,以实现科技理性与司法理性的有机统一,确保量刑公正的最终达致。

参考文献

- [1] 王熠珏. 论人工智能时代量刑规范化的实体进路[J]. 财经法学, 2025(3): 148-164.
- [2] 丰怡凯. 人工智能辅助量刑场景下的程序正义反思与重塑[J]. 现代法学, 2023, 45(6): 98-117.
- [3] 张玉洁. 人工智能辅助量刑的法律风险与制度建构[J]. 学术交流, 2023(3): 60-70.

-
- [4] 甄航. 人工智能介入量刑机制: 困境、定位与解构[J]. 重庆大学学报(社会科学版), 2023, 29(4): 191-202.
- [5] 王燕玲. 人工智能法律应用的发展瓶颈及方法论应对[J]. 政法论丛, 2025(2): 141-155.
- [6] 孙道萃. 人工智能辅助精准预测量刑的中国境遇——以认罪认罚案件为适用场域[J]. 暨南学报(哲学社会科学版), 2020, 42(12): 64-78.
- [7] 冯文杰. 人工智能辅助量刑公正取向的双重构建[J]. 华东理工大学学报(社会科学版), 2020, 35(6): 114-128.
- [8] 黄春燕. 法官量刑的自由裁量权与量刑公正的实现——兼论人工智能在量刑中的定位与边界[J]. 山东师范大学学报(社会科学版), 2021, 66(3): 136-145.
- [9] 孙道萃. 人工智能辅助定罪的进展、理论与应用[J]. 华南师范大学学报(社会科学版), 2024(2): 117-137+207.
- [10] 孙道萃. 人工智能辅助量刑的实践回视与理论供给[J]. 学术界, 2023(3): 112-128.
- [11] 张玉洁. 智能量刑算法的司法适用: 逻辑、难题与程序法回应[J]. 东方法学, 2021(3): 187-200.
- [12] 李立丰, 齐弋博. 数字算法时代的量刑公正及其构建[J]. 学术探索, 2022(4): 88-96.
- [13] 郑海味, 赖利娜, 章奕宁. 人工智能量刑系统应用的公正性、伦理价值及限度[J]. 治理研究, 2024, 40(3): 142-156+160.
- [14] 倪震. 量刑改革中“机械正义”之纠正——兼论人工智能运用的边界及前景[J]. 江西社会科学, 2018, 38(2): 192-200.
- [15] 王燕玲. 论大数据精准量刑对司法理性的精准描述[J]. 西南民族大学学报(人文社会科学版), 2021, 42(10): 103-113.
- [16] 张燕龙. 人工智能时代量刑的价值追求与挑战[J]. 中国政法大学学报, 2022(2): 264-276.
- [17] 洪涛. 正当程序视角下人工智能辅助量刑的挑战与重塑[J]. 东北大学学报(社会科学版), 2025, 27(2): 111-120.
- [18] 程龙. 人工智能辅助量刑的问题与出路[J]. 西北大学学报(哲学社会科学版), 2021, 51(6): 163-174.
- [19] 倪震. 量刑改革时代人工智能泛化之批判[J]. 人民论坛·学术前沿, 2020(7): 92-95.
- [20] 甄航. 智慧量刑的主体冲突、技术供给与方法回应[J]. 江苏社会科学, 2024(2): 175-184.
- [21] 崔仕绣. 智慧司法场域下量刑辅助系统之功能辨正[J]. 新疆大学学报(哲学社会科学版), 2022, 50(4): 30-38.
- [22] 李立丰, 齐弋博. 智能量刑算法决策与刑事自由裁量权的差异与融合[J]. 科技与法律(中英文), 2022(1): 10-17.