Advances in Social Sciences 社会科学前沿, 2025, 14(6), 346-356
Published Online June 2025 in Hans. https://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2025.146513

Hans汉斯

生成式AI赋能刑事量刑的法理反思、应用瓶颈及功能突破

卢　晗

新安威尔士大学法学院，澳大利亚　悉尼

收稿日期：2025年4月24日；录用日期：2025年6月9日；发布日期：2025年6月19日

摘　要

随着生成式人工智能(AI)技术的迅速发展，刑事量刑领域的传统模式面临前所未有的变革与挑战。本研究探讨了生成式AI在刑事量刑中的应用及其法理反思，重点分析技术引入过程中的三个阶段：排斥期、审慎期和共生期。研究指出，在排斥期，司法系统因合法性真空和技术黑箱特性对AI持否定态度；在审慎期，AI被视为法官裁量的辅助工具，但仍面临数据偏见与角色错位的挑战；进入共生期后，AI与法官的协同决策逐步深化，但同时也引发了可解释性鸿沟与伦理失控等问题。为了突破这些瓶颈，研究提出了多项阶段性功能突破方案，强调通过构建沙盒验证机制、责任矩阵、双轨意见制度等措施，在技术和法律之间实现动态平衡，推动“可解释——可问责——可协同”的新型量刑流程的形成。本研究最终认为，只有通过细致的制度与技术协同创新，生成式AI才能有效服务于司法公正，推动司法系统向智能化、公正化和透明化的方向发展。

关键词

生成式AI，刑事量刑，程序正义，技术黑箱，数据偏见，伦理边界

Jurisprudential Reflections, Application Bottlenecks, and Functional Breakthroughs of Generative AI-Empowered Criminal Sentencing

Han Lu

Faculty of Law, The University of New South Wales, Sydney, Australia

Received: Apr. 24th, 2025; accepted: Jun. 9th, 2025; published: Jun. 19th, 2025

文章引用: 卢晗. 生成式 AI 赋能刑事量刑的法理反思、应用瓶颈及功能突破[J]. 社会科学前沿, 2025, 14(6): 346-356.
DOI: 10.12677/ass.2025.146513

Abstract

With the rapid advancement of generative artificial intelligence (AI) technology, the traditional paradigm of criminal sentencing is undergoing unprecedented transformation and facing significant challenges. This study explores the application of generative AI in criminal sentencing, offering a jurisprudential critique while focusing on three distinct phases of technological integration: the phase of rejection, the phase of prudence, and the phase of symbiosis. The research observes that during the rejection phase, the judicial system adopts a negative stance towards AI due to a legitimacy vacuum and the opaque nature of algorithmic "black boxes." In the phase of prudence, AI is introduced as an auxiliary tool to assist judicial discretion, yet struggles with data bias and misalignment of institutional roles. As the symbiotic phase emerges, the collaborative decision-making between AI and judges deepens, but new dilemmas also arise, including an explanatory gap and ethical lapses. To overcome these bottlenecks, the study proposes several staged functional innovations, such as the establishment of regulatory sandbox mechanisms, accountability matrices, and a dual-track opinion system. These measures aim to foster a dynamic balance between technology and law, facilitating the evolution of sentencing procedures towards explainability, accountability, and human-AI collaboration. Ultimately, the study contends that only through meticulous institutional and technological co-innovation can generative AI genuinely serve judicial justice and propel the legal system towards greater intelligence, fairness, and transparency.

Keywords

Generative AI, Criminal Sentencing, Procedural Justice, Algorithmic Black Box, Data Bias, Ethical Boundaries

1. 问题的缘起

伴随 ChatGPT、文心一言等生成式 AI 的集中涌现，司法裁判这一传统“人类专属”场域被迅速推向数据–算法共同驱动的新拐点。尤其在刑事量刑环节，技术扩张与规范焦虑几乎是同时到达：一方面，生成式 AI 依托大模型的语义理解与内容生成能力，声称能够“读懂”卷宗、“复刻”法官经验甚至“预判”裁判趋势；另一方面，司法共同体却对其是否会侵蚀自由裁量、加剧黑箱化、冲击程序正当性而深感忧虑。长期以来，类似案件判决结果差异过大被视为损害司法公信力的顽疾，而统一裁判标准的缺失与法官主观性被普遍认为是“同案不同判”的直接诱因。对此，一种颇具吸引力的方案是借助算法“锁死”自由裁量空间：人工智能量刑系统因其“技术规范性”而非纯粹“计算预测性”被引入实践，其“封闭性与稳定性”被寄望于排除人为因素，输出“相对统一”的量刑结果[1][1]。对“同案不同判”症候的持续关注，为生成式 AI 闯入量刑领域提供了最初的正当化契机。生成式 AI 的语言生成能力为量刑系统带来两个显著变数：其一，“理由生成”不再只是呈现结构化要素，而是能够输出类人文本，甚至主动解释量刑逻辑；其二，模型参数远超传统算法，训练语料来源广泛，潜藏的偏见、歧视与安全风险更难被观测与校验。这种“解释力提升”与“黑箱加深”的并存使原本就存在的合法性、透明度与问责挑战进一步复杂化。于是，“允许 AI 走多远”与“人对 AI 说最后一句话”的攻守边界重新勾勒。回望技术推

[1]人工智能量刑系统的‘封闭性’和‘稳定性’能够排除主观裁量，进而导出相对统一的司法结果。

动的初衷，人工智能量刑希望通过数据–算法方法论减少偏差、提高效率、统一标准；回看规范层面的焦虑，则聚焦于程序正义、法官主体性与被追诉人权利能否获得同频保护。正是在“促进量刑公正”与“防止算法侵蚀”之间来回摆荡的张力，构成了生成式 AI 赋能量刑的时代命题[2][2]。本研究据此提出：只有通过细致辨析技术优势与法律底线的交汇区，并以分阶段、分层次的治理工具重塑“可解释——可问责——可协同”的新型量刑流程，才能在算法与法治之间建立动态平衡，让技术真正服务于司法公正与公众信赖[3]。

2. 法理反思：生成式 AI 被刑事量刑“接纳”的三阶段演进

2.1. 排斥期：公权独占与风险防御下的拒斥逻辑

生成式 AI 初露锋芒之际，刑事司法体系的第一反应并非欢迎，而是本能排斥。这种拒斥既源于对司法主权的坚守，也折射了对技术不确定性的风险防御。首先，量刑历来被视为“公权力最核心的自由裁量区”，法官得以在罪刑均衡、个别化处罚之间作纵横取舍。倘若让算法介入，便意谓着国家垄断裁量权的“外流”，于是出现“算法不可裁判”“技术威胁司法主权”等论调。并且，传统程序正义理论以物理诉讼空间为场域，强调当事人“在场”“对质”，而生成式 AI 的训练、推理与输出却全部发生在数字空间，使得“信息可感知”难以转化为“信息可解释”进一步加剧制度排斥感。在既往技术治理经验中“自动化决策黑箱”“偏见数据毒化”已多次引发公共安全与权益侵害，刑事司法尤其担忧此类风险在量刑环节“放大”。于是，各级法院、检察机关对生成式 AI 采取消极审慎态度，常以“技术尚不成熟”“伦理评估缺位”为由拒之门外，排斥期由此形成[4]。

2.2. 审慎期：工具理性支配与条件接纳的双轨逻辑

随着案例数据库扩容、“23 + 2”智能辅助平台等项目示范成效渐显，“AI 可用于量刑”的可行性获得学界与实务界阶段性共识。进入审慎期，司法系统一方面承认算法的工具价值，另一方面又设置多重“保险栓”。这种双轨逻辑主要表现在：其一，程序功能定位的“有限介入”。智能量刑辅助技术被明确定位为“规范量刑裁量权的工具”，其核心任务是通过类案数据“锚定”量刑区间，防止主观恣意，而非直接取代法官自由裁量。正如研究所言，“人工智能以类案量刑信息大数据为蓝本，将共同规律注入个案量刑之中，从而约束可能出现的偏差”[5]。其二，程序运行模式的“人主机辅”。最高人民法院《人工智能司法应用意见》确立“辅助审判原则”，要求坚持“法官主导、AI 协同”的责任矩阵，防止“技术统治”滑向“技术依赖”。其三，程序正当性的“前置过滤”。在实践中，越来越多的法院引入“算法合规评估”“模型偏见检测”等前置性程序；同时，对当事人的参与权、质疑权进行原则性赋权，例如允许控辩双方就类案库选取、要素抽取方式发表意见。然而，由于缺乏成文的量刑算法正当程序规范，上述保障仍呈碎片化，出现“原则宣示有余、操作细节不足”的裁判落差。

2.3. 共生期：协同自治与价值重塑的融合逻辑

“共生期”标志着生成式 AI 在刑事量刑领域从一种外部植入的“异体”演变为与司法系统深度嵌合、相互塑造的“共生体”。此阶段的问题核心不再是排斥或审慎接纳，而是如何在这种新型的、具有潜在“智能涌现”特征的人机关系中，确保 AI 的“自治性”增长始终被“可信性”框架所约束，并将抽象的法律价值与伦理原则转化为可计算、可执行、可验证的技术内在属性。这要求我们超越“人机分工”的静态视角，转向一种“协同进化”的动态治理逻辑，其中技术的发展路径、法官的认知模式以及司法流程本身都在持续的互动中被重新定义。

[2]关于传统程序正义理论主要要素及其对数字诉讼空间的适用局限。

所谓的“AI 预判–法官复核–理由共写”闭环模式其深层意义在于构建一种“交互式认知增强”而非简单的“决策外包”。AI 的“预判”不应被视为标准答案，而是一个基于大规模数据模式识别的“高维参照系”，它能揭示人类法官可能忽略的潜在关联或一致性偏差。法官的“复核”则成为关键的“价值锚定”环节，运用其法律专业知识、经验直觉和对个案特殊性的把握，对 AI 的概率性输出进行规范性校验与情境化调适。而“理由共写”的理想状态并非简单的文本拼接，而是通过结构化的论证界面(如基于图尔敏模型的可视化论证工具)，促使法官与 AI 就关键量刑因子、证据权重、法律适用等进行“对话式”推敲，AI 负责提供数据支持、逻辑一致性检查、备选论证路径，法官则主导价值判断、理由选择与最终表述，形成一份既有数据支撑又体现司法裁量温度的“增强型判决理由”。

3. 应用瓶颈：生成式 AI 被刑事量刑阶段性困境的多维剖析

3.1. 排斥期的双重迷惘——缺规与黑箱

“排斥期”实质上是一种心态与制度层面的抗拒状态。这种抗拒源于“双重迷惘”：“缺规”(规范缺失)与“黑箱”(技术不透明)。这两种迷惘相互交织、互为因果，共同构筑 AI 进入司法核心地带——量刑——的高墙。

3.1.1. 合法性真空：缺乏前置评估导致引入门槛失衡

要真正把握“合法性真空”的根源，应回到司法权的宪制定位与数字技术的制度属性之间的张力。司法裁判的独占性来自宪法意义上的“权威性—可复议性—可预测性”三要件，而生成式 AI 的本质是“概率性–自学习性–不确定性”。二者发生耦合，必然需要一个“范式转译”机制，把 AI 的技术语境转译为司法领域能够理解并控制的规范语境；然而当下中国(乃至全球)正缺少这一层“翻译规程”。欧盟《AI Act》提出“高风险 AI”需进行“合规性评估 +CE 标识”，其核心思想是以行政前置许可替代事后侵权救济；美国则在《Algorithmic Accountability Act》中强调“影响评估 + 外部审计”的双轨模式。中国虽然已发布《新一代人工智能治理伦理规范》《自动化决策个人信息保护规定(草案)》等软法，但针对刑事实体裁判这一“高耦合、高风险”场景仍无硬法框架[6]，导致地方法院往往直接把商业化“类案检索”系统平移为量刑辅助工具，忽视了刑事审判对证据严谨度和程序正当性的高阶需求。

更深层的问题在于“合法性生产过程”司法机关缺乏一套把技术风险转化为可管理节点的评估链条。如“算法影响力评估”(AIA)、“数据保护影响评估”(DPIA)原应成为入场门槛，却被简单等同于第三方测试报告；“司法适用性评估”应当检验模型在不同罪名、不同地域、不同群体上的偏倚分布，却往往由开发商自填问卷了事。结果就是“低质算法”“快车道”入场，优质算法因评估机制缺位也拿不到“准入通行证”，形成所谓“门槛失衡”。要化解该空窗，需要引入“多层次—多主体—多维指标”的立体评估矩阵：一方面通过行业标准把模型精度、偏见度、可解释度量化为可比指标；另一方面设立“司法算法认证”独立机构，对每一次版本迭代进行动态溯源与再认证，并将认证结果与司法采购、预算拨付、责任保险挂钩，形成“合法性–经济激励”闭环。

3.1.2. 责任模糊：技术黑箱引发归责断层

在黑箱之所以引发归责断层，关键不在“看不见”，而在“剥离不出因果链”。传统侵权法的因果认定依赖“事实推定 + 经验法则”：假如一辆汽车制动失灵导致事故，刹车系统设计缺陷可被经验推定为主要原因。而深度学习模型牵涉的是高维参数空间与动态权重更新，其决策路径呈“多因素交互”结构，一处细小权重扰动即可导致结果巨变，无法用线性因果模型解释。此时，“谁的过错”就转化为“谁应该承担风险”，责任分配的逻辑需从“归责型”转向“风险管理型”。欧盟《AI 民事责任指令草案》

提出“可推定因果关系”，只要受害人证明系统不符合《AI Act》规定，即可推定开发者或部署者承担赔偿责任；此种“合规背书–责任豁免”模式为中国立法提供了镜鉴：在刑事AI场景，可考虑将“白盒可审计 + 日志留痕”作为开发者的免责前提，把“偏见测试 + 模型冻结”作为部署者的持续义务，并以“显失合理”标准检验法官对AI输出的依赖程度——当法官未对显著异常输出进行复核即采纳时，可推定其存在过失。进一步地，归责断层还体现为“多主体共享信息的不对称”，开发者掌握算法秘密，部署者掌控系统更新，法官掌有输入数据，却没有任何一方单独握有完整链条。对此可通过“分层日志 + 分布式签名”把责任链数字化：模型侧记录参数与梯度变动、部署侧记录版本与更新时间、使用侧记录输入输出与人工调整，并用区块链或可信执行环境进行跨主体哈希对齐；一旦出现错误量刑，可在数秒内回溯到“责任裂缝”所在的具体时间戳和责任人。只有当“技术不透明”与“责任可定位”实现对冲，司法系统才能在面对AI失误时迅速修复信任裂口。

3.2. 审慎期的结构桎梏——数据与角色

“审慎期”意味着保留性接纳，通常伴随着“人主机辅”的基本原则框架。在这种框架下：“数据”的先天缺陷与“角色”的实践错位。这不再是能不能用的问题，而是用了之后如何确保用得好、用得对的问题。

3.2.1. 数据偏见：训练语料失衡致量刑结果系统性歧视

要把“数据偏见”真正拆解透，必须同时从“符号层–制度层–社会层”三重视角介入。符号层面，司法文本天然具有“结构化–去情境化”的书写规则：判决书以罪名、法条、量刑情节为主干，弱化了被告人阶层处境、社区支持度等“软数据”，这使得模型在编码时就丢失了衡量“社会脆弱性”的变量；当这类变量与重刑倾向高度相关时，算法会把它们“隐性映射”为诸如“前科次数”“户籍类型”等可量化特征，在数学空间完成对弱势群体的重新标记。制度层面，则是“案例采集–审级筛落–文本脱敏”环节的选择性漏斗：基层法院高发、轻罪量刑短、裁判文书上网率高，反而让“低端犯罪–贫困群体”在训练集占比畸高；而重大、敏感、涉权势案件常因“社会影响考虑”被剥离公开渠道，使训练集几乎看不到“白领犯罪—缓刑/短刑”的真实基线，机器遂误把“弱势–应重刑”当作数据真理。社会层面，则表现为宏观结构的不平等沉淀——区域经济差距、执法资源差异、侦查取证能力不均衡在历史数据中烙下深痕。法律语言把这些差异“合法化”地编码为“案情证据充分度”“被告人认罪态度”等量刑因子，算法在无监督聚类中自然把它们归入“重刑簇”，最终形成对贫困地区或特定族群的系统性负面标签[4]。

这种偏见之所以难以察觉，是因为司法AI评价体系传统上仅盯“准确率”“一致率”，而不考察“差异性分布”。要揭露并矫正就须引入“分层公平指标”与“反事实对比框架”：对每一敏感属性建立对照组，用反事实方法模拟“同案异人”与“同人异案”双重场景，量化偏见边际；再辅以因果推断，截断“社会脆弱性→执法资源→证据完备度→量刑”这一链式因果中的桥变量，通过调整权重或加设规则，使模型区分“合法差异”与“结构性差异”，将后者从量刑预测空间剥离。

3.2.2. 权责漂移：人机角色错位弱化法官裁量主体性

“权责漂移”绝非单纯认知依赖，更是一种“组织理性–技术理性–法律理性”冲突下的结构现象。组织理性层面，法院绩效考核强调结案数量、平均审理周期、类案同判率，使法官面临“低冲突–高吞吐”的隐性激励，AI所谓的“默认建议”天然契合这种指标偏好；技术理性层面，深度模型的“概率输出 + 置信区间”与司法的“确定判决 + 确信标准”逻辑错位，法官在缺乏概率思维训练时更愿意接受系统给出的“区间中位数”作为安全锚点；法律理性层面，刑罚决断需要情节加权与价值平衡，但AI训

练目标仅是最小化历史误差—这一“经验主义法则”与刑法的“规范主义法则”发生错配，导致机器意见在形式上看似合规，实则缺少规范演绎支撑，却仍被法官误当作“经验共识”。

漂移链条的关键接口是“人机交互设计”。当前多数量刑辅助系统界面先展示机器区间，再让法官手动调整；这一流程本身就对人类施加了“算法先验”压力。当法官“轻触鼠标”即可结束决策，而“越线修正”却要逐条理由输入、层报复核，制度性摩擦成本让“顺从算法”成为理性选择。更进一步，判决生成系统常把 AI 建议直接嵌入量刑说理模板，法官稍作编辑即能出具完整文书，这种“文书拼贴”让人机实质分工模糊，责任链由此断裂。要重新锚定法官主体性，关键在“决策权—解释权—风险承受权”三权合一：决策权方面，界面设计应改为“盲审式”——先录入人类意见，后揭示模型建议；解释权方面，系统必须提供可核查的“差异报告”，使法官对每一数值差异都有备可查；风险承受权方面，则需引入“算法采信险 + 责任反推金”机制——法院对完全采纳 AI 建议的案件购买第三方责任险，一旦发生过刑，保险公司凭系统日志追偿开发商；开发商为了降低索赔风险，反向激励其提升模型解释度与抗偏差能力。从而把裁量主体性、说理义务与风险负担重新捆绑回法官本位，使 AI 真正退回到“智能书记员”而非“隐形法官”的角色[7]。

3.3. 共生期的深层挑战——解释与伦理

3.3.1. 可解释性鸿沟：生成式输出难以满足说理义务

从知识论维度看，人类量刑说理是“规范理由 + 事实叙事”双层结构：先抽取法律规范的要素框架，再在其中填入对具体事实的评价和价值衡量；生成式模型则遵循“语料分布–概率采样–语义拼接”的统计机制，两者在生成动因与评价标准上呈现根本错位[8]。具体表现为(表 1)：

Table 1. Comparative table of risk mechanisms and legal implications of generative AI-assisted sentencing reasoning
表 1. 生成式 AI 辅助量刑说理的风险机制与法律影响对照表

风险类别	技术表现/机制	法律/司法核心要求	问题机制与细化风险	潜在系统危害
语义外观与规范内核的剥离	仅捕捉文本共现概率，重现标准句法模板(如“先法条、后情节”结构)	判决理由需逻辑推演、事实与规范紧密牵连	要素罗列合乎表面逻辑，缺乏规范逻辑和价值匹配，内在张力断裂，言之无物	说理空洞，失去个案公正与权威，公众难以信服
解释功能与预测功能的冲突	为流畅性/准确率优化“语言熵”，生成平均化、无争议的文本	司法说理应凸显事实争议、权衡纠结	案件独特矛盾被抹平，判决趋于模板化，争议焦点难以呈现	判决理由趋同、解纷功能弱化、失去个案温度
“事后合理化”陷阱的扩散	SHAP/LIME 等可解释性工具展示统计权重，简单注释替代理解和三段论	说理需揭示因果性、规范关联、裁量逻辑	仅有变量权重、缺乏规范–事实–价值三层衔接，法官易依赖幻觉解释，主动推理缺失	判决理由批量化、推理虚化、责任虚化
法治可预测性的被稀释	不同案件同模型生成文本高度相似，细微参数波动致隐含规则难以归纳	法律确定性与“同案同判”、可预测性	长文本亦无稳定的裁判规则，个体差异莫名，法律共同体缺失参照	“同案不同判”、“规则漂移”、法治权威受损

3.3.2. 伦理边界失控：算法自治突破程序性公正的底线

算法自治在共生期表现为“三重替代”：替代感知、替代判断、替代裁量。每一层替代都对程序正义构成新的撕裂点。第一，在感知替代层——证据筛选与事实摘要由模型预处理导致案件材料输入呈“算法预成像”；辩护方面对的并非完整卷宗，而是机器先行挑选过的“显著特征向量”。此举在源头上重塑“争点框架”[4]，被告人实际上被排除在证据完整性审核之外，违反对抗制“对等呈现证据”原则。第二，在判断替代层——模型在量刑区间预测时引入“风险标签”逻辑(如再犯概率、社会危害度分级)，

而这些标签源于历史统计而非个案具体情节，易把宏观治安治理目标嫁接到个体刑罚决策中，形成“群体治理先于个体正义”的治理错位。此时，刑事司法悄然服从于治安算法统治，而非宪法层面的“罪刑法定–责任自负”。第三，在裁量替代层——生成式 AI 与流程自动化结合可直接输出“量刑建议 + 生成理由 + 附条件释明”，法官只需在系统界面点击确认即可。长此以往，量刑自由裁量被算法“默会俘获”，而算法开发者则在系统参数中暗中置入政策权衡乃至商业考量，实现“规则外包”。这样形成的“技术路径依赖”具有不可逆特征：一旦法院、人事、预算体系围绕软件重构，退出成本指数级飙升，导致纠错机制形同虚设。

4. 功能突破：阶段对应的制度与技术解法

4.1. 排斥期的破局——从“风险否定”到“程序性容纳”

针对排斥期因“合法性真空”和“责任模糊”而导致的技术拒斥，破局之道在于构建早期介入的程序性机制，将对技术的否定转变为有条件的程序性容纳，从而为后续的探索与发展奠定基础。

4.1.1. 沙盒验证 + 动态评估：为合法性开辟前置通道

创设并整合“沙盒验证”的前置过滤与“动态评估”的持续监管方案，该机制的核心价值在于构建一个动态的、嵌入式的、生产性的、治理性的框架。这标志着对待新兴司法技术(尤其是高风险 AI)的态度，从过去简单粗暴的“要么放任自由发展、要么彻底禁止应用”的二元极端思维，转向一种更为成熟和精细化的程序性容纳逻辑。“程序性容纳”的核心在于不预设最终的“允许”或“禁止”，而是设计一套严格的程序。技术的命运取决于其能否持续满足这套程序所设定的标准。这套程序依赖于过程控制(如沙盒阶段性测试)、证据积累(如性能数据、伦理报告)和适应性调整(如基于评估结果更新模型、调整应用范围或撤销许可)。技术在满足条件的前提下被逐步、审慎“容纳”进司法体系，同时保留了在风险失控时进行限制或退出的机制。

“风险分级”沙盒的核心驱动力不是技术复杂性，而是伦理风险的严重性(具体体现为对公民自由权(刑罚轻重)的潜在影响程度)。在司法领域，技术的最大风险在于对基本人权(尤其是自由权)的潜在侵害。因此，风险评估必须以人权影响为首要考量，而非技术本身的先进或复杂程度。对公民自由权影响越小的场景(如极轻微罪行、程序性辅助)，准入门槛和测试要求相对较低；影响越大的场景(如重罪量刑建议)，准入门槛和测试要求指数级提高。这构成了伦理风险梯度。沙盒通过递进式测试(从低自由刑到重大刑罚案件)，在技术内在的不确定性与司法判决后果的严重性之间，不断探寻和校准一个动态平衡点。为此应遵循比例原则与预防原则，前者侧重于技术应用的范围、强度、介入程度，必须与其可能带来的风险和要解决的问题相称。不能为了微小的效率提升而承担巨大的公正风险。后者旨在关注风险存在科学不确定性，但有理由相信可能造成严重损害时采取预防措施。宁可牺牲部分效率，也要优先防范对核心权利的侵害。

彻底颠覆“一次审批定终身”的静态许可模式，将合法性视为一个动态的、需要在使用过程中持续生成和不断再确认的过程。AI 模型性能受到数据漂移、现实情况变化导致输入数据分布改变、算法更新、模型自身迭代、使用环境变化、部署场景、用户习惯改变等多种因素影响，其有效性和安全性并非恒定不变。AI 系统的合法性基础从过去单一依赖前置审批的“准入许可”(如同拿到一张“准生证”)，扩展为一个包含性能表现(Performance)、伦理合规(Ethics)、用户信任(Trust)在内的多元、动态指标体系。最终目标是确保 AI 辅助量刑系统在其整个生命周期内，始终处于持续、有效的监管之下，真正实现从“准生证”到“健康证 + 年检报告”的治理模式转变。

4.1.2. 责任矩阵 + 多元治理：厘清“技术开发 - 部署 - 使用”三方义务

责任矩阵的核心是把抽象的“最终由人负责”拆解成“谁对什么负怎样的责任”。技术开发方被置于“产品安全 + 算法质量”双重责任中心：需向监管机构递交“模型说明书”(涵盖数据来源、特征工程、偏见测试、可解释性方案与更新日志)，并接受“白盒验证 + 密钥托管”的双轨检查——监管部门在保存知识产权机密的前提下可进入“受控白盒”查看源代码与权重层级，实现真正的可审计性。系统部署方(法院或其技术部门)承担“场景适配 + 基础设施安全”责任：首先要完成“司法采购尽责调查”，对供应商过往合规纪录、模型版本、恶意篡改风险进行尽调；其次要在本地搭建“算法使用日志 + 决策链溯源库”，确保每一次量刑建议的输入、输出与人工调整痕迹均可回溯。最终使用者——法官——则被赋予“专业判断守门人”责任：须进行“算法证书”培训并通过考核，才能启用系统；在判决书中需写入“算法参考说明”板块，说明 AI 建议采纳或拒绝的理由；若选择完全采纳 AI 建议，则需进入“二次审签”程序，由资深法官复核。矩阵外层再套上一张“多元治理网”：技术伦理委员会常驻成员由开发商代表、法院人员、律师协会、数据科学家及社会公众代表组成，定期公开发布“风险观察清单”；律师和公益机构可启动“算法质询程序”，对疑似偏见实例提起听证；行业协会制定的“司法 AI 操作规范”作为软法，与强制性法律条款形成“硬软嵌套”，通过多元主体持续交互来填平归责断层，把责任落到最真实、最细粒度的操作层面。

4.2. 审慎期的升级——重塑“人机共治”量刑流程

4.2.1. “双轨意见 + 最终理由”机制：巩固法官主导权与说理义务

“双轨意见 + 最终理由”机制的价值不在于提供“另一个选项”[9]，而在于通过特定的流程设计改变裁判的“生产方式”。它拒绝将 AI 视为与法官平行的“决策者”，而是将其定位为在特定节点、以特定方式介入的“受控辅助者”。面对 AI 可能带来的“自动化偏见”或“责任稀释”风险，该机制通过强制性的程序步骤，确保最终裁判权牢牢掌握在法官手中，并且这种掌握不是形式上的，而是体现在独立的心证形成过程和承担充分说明理由的责任上，强调法官不仅是“拍板者”，更是思考过程的“主导者”和“责任人”。即使 AI 建议非常精准，如果法官未能独立思考、理解并认同其逻辑，裁判的合法性和正当性也会受损。因此，机制设计的重心在于确保过程的独立性、透明性和规范性上，使法官的思考过程本身成为可被追溯、可被审视、可被问责的对象。

运用“延迟揭示(Delayed Disclosure)”与“并行输入(Parallel Input)”策略要求法官在看到 AI 意见之前，必须先行独立录入自己的初步量刑区间(或其他关键判断)与主要理由。这两个动作在逻辑上是“并行”发生，但法官的输入是前置的。同时，回应认知心理学中关于“锚定效应[3] (Anchoring Effect)”的风险。若 AI 建议先入为主，即使法官主观上想保持独立，其潜意识可能受其影响，导致思考范围受限或将精力用于“证伪/证实”AI 建议，而非基于事实和法律进行原生性思考(Original Deliberation)。这本质上是对法官认知能力的尊重，而非贬低。必须承认即使是经验丰富的法官也可能受到认知偏误的影响，因此需要制度性的保障。通过程序设计从源头上强化“法官自主判断”的实质内涵。通过时间戳、修改痕迹、引用资料与案例的完整留存，记录从法官初步意见、参考 AI 意见、到最终裁判理由形成的完整轨迹[10]。揭示法官的信息检索范围、参考先例的选择、心证变化的可能路径，为理解裁判的深层逻辑提供依据。为上级法院复核、审判管理监督、法学研究提供前所未有的透明度和深度。这种透明性是程序正义的重要面向，使裁判过程不再是封闭的“黑箱”，增强司法活动的可信度。

4.2.2. “数据全周期治理 + 去偏见审核”体系：提升量刑数据质量与可信度

全周期治理的起点是“数据准入注册制”——任何进入训练池的司法案例先进入“元数据登记簿”，

[3]锚定效应指个体在决策时，会过度依赖接收到的第一个信息(锚点)，后续判断会围绕这个锚点进行调整。

记录来源法院、审理层级、案由标签、涉敏感特征标记和脱敏状态，实现数据血缘可追溯。在清洗阶段建立“多维偏差探测面板”，按照地域、性别、族群、量刑幅度、辩护类型等维度生成分布热图，自动比对全国基准分布与历史趋势，异常波动触发“人工抽检 + 专家复核”双通道处置。标注环节推行“OCR + 法律语义本体”技术，确保罪名、情节、量刑因子被同义对齐，减少因表述不一引起的“语义漂移”偏误。训练阶段则引入“公平性共识指标池”：在群体公平(Statistical Parity)、条件公平(Equalized Odds)、个体公平(Counterfactual Fairness)三类指标中，由法院、检察、律师、数据科学家共同选取底线指标，为模型调参设定约束；配合对抗性去偏算法，实现“公平指标–准确率”双目标优化。上线后运行“偏见哨兵”：系统实时输出后按敏感变量聚合生成“偏见雷达图”，若某一群体量刑显著偏重，自动进入“回滚实验”，即以同案事实在模拟环境重新跑模型并与当下版本对比，确认是模型漂移还是数据输入问题。外部监督方面，设立“司法数据信托”模式——由第三方数据信托机构托管经脱敏的训练集与更新集，诉讼当事人和研究者可在受控沙箱内调用数据进行独立检验；同时为被追诉人设置“数据质疑权”快速通道，可就其个人或相似群体数据提出偏见指控，并触发信托机构的强制偏见复审。通过技术与制度双轨运作，将数据问题从“隐形风险”转化为“可监测、可纠正”的透明流程，使量刑算法的每一次输出都建立在动态更新且经多方审核的可信数据之上[9]。

4.3. 共生期的深化——迈向“可信自治”与价值嵌入

当生成式 AI 能力进一步提升，迈向更高程度的“自治”时，应对“可解释性鸿沟”和“伦理边界失控”挑战，需要将信任机制内置于技术本身，并将价值原则深度嵌入算法设计。

4.3.1. 全链条可解释协议 + 算法水印：实现透明、可溯与可问责

面对生成式模型数以十亿计的参数与非线性耦合，仅凭“代码公开”远不足以满足司法场景的可解释需求，因其既缺乏语义可读性，无法复现实时推理环境。全链条可解释协议应当对“数据输入–特征提取–语义生成–输出呈现”四级流程逐一定义强制性可记录字段与可追溯格式：在输入层，以“概念标引 + 哈希指纹”方式固定卷宗要素；在特征层，利用可视化深度特征图与概念激活向量(TCAV)捕捉哪些法律要素被显著关注；在语义生成层，要求模型实时产生“梯度归因–SHAP 权重–因果推理图”三类解释元数据并写入不可篡改日志；在输出层，则用“结构化理由模板 + 自然语言摘要”双轨展示，使法官既能获得高度概括的说理，也能随时下钻查看技术细节。算法水印技术进一步解决“谁写的理由”与“理由被改过吗”的问题：通过在隐向量空间注入低幅度、具有法律机构唯一标识符的不可见签名，实现对生成文本的鲁棒溯源；再配合零知识证明与区块链时间戳，把模型版本号、参数哈希、训练批次 ID 与具体输出之间建立不可伪造的链式关系。如此一来，任何人若试图删改 AI 生成的理由或在未授权的模型上复用理由，都会在验证端暴露。该机制不仅满足《欧盟 AI 责任指令草案》中“自动化决策的可审计”要求，也为上诉法院提供“一键还原”判决生成过程的技术凭据，真正把技术黑箱转化为可校验、可反驳的“灰盒”体系。

4.3.2. 伦理算法嵌入 + 自我纠偏模块：确保自动化演进符合正当程序与实体公正

“伦理算法嵌入 + 自我纠偏模块”旨在通过将法律与伦理原则深度融入算法设计，并赋予 AI 系统动态自我修正的能力，实现司法人工智能从被动遵循外部规则向主动内化价值规范、并确保持续合规的范式跃迁。在算法“基因层面”(目标函数、参数空间、推理逻辑)直接编码法律与伦理的核心要义，构建一种具有“计算性道德感”与“程序性自省力”的智能系统。考虑到，司法裁判的历史数据是过去的反映，可能包含了已被社会和法律所摒弃的偏见。单纯拟合历史，可能导致 AI“学会”并固化这些偏见。同时，对于法律原则未明确覆盖的新情况，缺乏价值引导的 AI 可能给出逻辑上“自洽”但法理上或社会

Table 2. Case analysis and application evaluation of "full-chain explainability protocol + algorithmic watermarking" (combined with Deepseek-explainable AI conception)
表 2. 案例分析与“全链条可解释协议 + 算法水印”应用评估(结合 Deepseek-Explainable AI 设想)

案例信息	罪名	裁判要旨核心	Deepseek-Explainable AI 的可解释性与可追溯性体现(基于“全链条可解释协议 + 算法水印”)
田某阳、沈某贤案(2021)鄂 9021 刑初 86 号	危害国家重点保护植物罪	环境资源案除考虑法益侵害外，还需考虑生态破坏程度及修复可能；主观恶性小、认罪认罚、积极修复的从宽	**数据输入层：**涉及生态破坏评估报告、修复情况证明、被告人认罪认罚具结书、罚金缴纳凭证等被标引。 **特征提取层：**TCAV 关注“生态环境修复程度(树苗成活率)”、“被告人主观恶性(是否为初犯偶犯)”、“认罪认罚态度”、“积极缴纳罚金”等。 **语义生成层：**梯度归因和 SHAP 权重清晰显示“树苗全部成活”、“自愿认罪认罚”、“全额缴纳罚金”等因素对“建议从宽处罚”的积极贡献 **输出呈现层：**结构化理由：“被告人田某阳、沈某贤擅自采挖国家重点保护植物 XX 株，但其行为主观恶性较小(证据：调查笔录链接)；案发后能自愿认罪认罚(证据：认罪认罚具结书链接)，积极缴纳罚金(证据：缴费凭证链接)；且涉案树苗经专业移栽已全部成活(证据：国家公园管理处证明链接)，生态环境得到最大限度修复。本模型认为符合从宽处罚条件” **算法水印：**保证理由的原始性和机构归属

观感上难以接受的结果(如表 2)。因此，单纯追求技术指标(如预测精度)在司法领域是危险且不充分的。

为此，利用“四维效用(Four-Dimensional Utility)”优化框架将预测精度、群体公平、个体公平、合法性符合度作为算法共同优化的目标。四个维度之间可能存在冲突(例如，提升群体公平可能轻微牺牲预测精度)。算法设计需要在这些维度之间进行显式的权衡(Trade-off)与整合(Integration)，找到一个符合司法价值的平衡点。不可否认的是，将高度抽象、依赖人类理解的法律原则转化为机器可理解、可执行的数学形式的关键一步，是 AI 伦理工程的核心难点之一。正则项在机器学习中通常用于防止模型过拟合(过度学习训练数据的细节和噪声)增加模型的泛化能力。在此处被赋予新的使命：引导模型在学习数据模式的同时，倾向于生成符合刑法比例原则(或其他法律原则)的量刑分布。具体而言，将纠偏逻辑置于“可信执行环境(Trusted Execution Environment, TEE)”中，确保这套内部治理机制本身不被篡改、失效或“做假账”。利用硬件安全特性(如特定的处理器区域)创建一个隔离的运行环境。在这个环境中运行的代码和处理的数据，其机密性(不被外部窥视)和完整性(不被外部篡改)受到硬件级别的保护，即使是操作系统或更高权限的软件也难以干扰。同时，设立外部监督 API (Application Programming Interface)，向授权的第三方(如技术伦理委员会、司法监督机构)暴露算法的实时“健康指标”。该接口允许授权的、独立的监督者实时或定期获取关于伦理算法和纠偏模块运行状况的触发探测器(检测到潜在偏见或违规)的频率、类型、具体情况，以及系统采取的响应措施及其执行结果等关键信息。当于为“算法内部监察院”(由伦理算法和纠偏模块构成)设立了一个外部审计机制。监督者可以通过 API 获取的数据，评估内部治理机制的有效性、是否存在系统性问题、是否需要进行干预或调整。

5. 结语

研究通过对生成式 AI 在刑事量刑领域应用的多维度剖析，得出如下结论：生成式 AI 在刑事量刑中的应用初期面临合法性真空、技术黑箱、数据偏见等问题，但随着技术的进步与制度的完善，其应用前景逐渐明确，呈现出从排斥到审慎、再到共生的三阶段演进过程。研究提出如沙盒验证、双轨意见机制、数据全生命周期治理等功能突破路径，能有效应对当前的技术困境，推动 AI 与法治的深度融合，确保技

术在提升量刑公正性的同时保证程序正义与司法透明。最终，提倡通过法律与技术的协同发展，并在伦理与透明度上进行严格规范，实现生成式 AI 在刑事量刑中的高效应用，并保障司法公正与公众信任。

参考文献

[1] 丁晓东. 人机交互决策下的智慧司法[J]. 法律科学(西北政法大学学报), 2023, 41(4): 58-68.

[2] 陈瑞华. 程序正义理论[M]. 第 2 版. 北京: 商务印书馆, 2022: 188-196, 251-256.

[3] 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.

[4] 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.

[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] 严驰. 论人工智能标准与法律协同治理体系之构建[J]. 北京理工大学学报(社会科学版), 2024, 26(6): 109-121.

[7] 塔妮娅·索丁, 王蕙, 李媛. 法官∨机器人: 人工智能与司法裁判[J]. 苏州大学学报(法学版), 2020, 7(4): 10-22.

[8] 郑海味, 赖利娜, 章奕宁. 生成式 AI 赋能刑事量刑的法理反思、应用瓶颈及功能突破[J]. 治理研究, 2024(3): 50.

[9] 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.

[10] 刘艳红, 徐珉川, 张柏礼. 构建立法可信生成智能推动创新赋能立法实践[J]. 中国法治, 2024(7): 20-27.