聚知刊出版社
JZK publishing

生成式人工智能辅助量刑的风险治理与实现路径

龙颖 胡兰之 王茹薪 胥兆伟

西南科技大学 法学院，四川绵阳，621000；

摘要：本文将以 Chat GPT 为人工智能应用的范例，采取实证研究法，从具体算法的技术特征出发，厘清生成式人工智能与司法裁判的融合风险，并以大数据与机器学习为技术依托，以自动生成量刑预测意见为目标导向，以为量刑裁判者提供参考、减少量刑偏差为价值取向，重新省思人工智能参与司法审判的限度。通过对人工智能司法审判的应用路径进行重新规划，探索人工智能的工具理性与司法审判的价值理性重归和谐的法律方案，从而达到在推进司法革新、提高司法效率的同时维护好司法权威性的终极目标。

关键词:人工智能；辅助量刑；司法权威；风险治理

DOI：10.69979/3029-2700.25.07.040

1 生成式人工智能的运行逻辑——以 Chat GPT 为分析对象

本节将以 Chat GPT 为分析对象，从技术架构、训练机制与生成过程三个维度，系统解析生成式人工智能的内在运行逻辑及其技术特征。

1.1 人工智能运作之底层技术架构

生成式人工智能的效能高度依赖数据的质量与多样性。Chat GPT 的语料库覆盖法律条文、裁判文书、学术论文等多源文本，既包含结构化数据（如罪名与刑期的对应关系），也涵盖非结构化数据（如案情描述与社会舆情）。在处理过程中，数据需经过清洗、去噪与标注，例如通过命名实体识别（NER）技术提取法律文书中的关键要素（如被告人信息、犯罪情节），并构建法律知识图谱（Legal Knowledge Graph）以增强语义关联性。

1.2 人工智能辅助量刑的技术原理

人工智能辅助量刑的核心在于通过大数据与算法模型模拟司法裁量逻辑，其技术实现路径可分为以下三个阶段：

1.2.1 法律知识图谱构建与数据挖掘

法律知识图谱是人工智能量刑的基础框架，其通过结构化处理法律法规、司法解释、裁判文书等非结构化数据，形成法律要素间的关联网络。以“小包公”系统为例，其聚合了超过 330 万部法律法规文件及 1.5 亿份裁判文书，并利用自然语言处理（NLP）技术提取案件中的定罪量刑要素，如犯罪情节、法定刑幅度、地域司法实践差异等，形成覆盖多维度法律关系的知识图谱。这一过程不仅解决了传统法律检索的碎片化问题，还为后续模型训练提供了标准化数据支持。

1.2.2 类案识别与机器学习模型训练

基于知识图谱，系统通过机器学习算法对海量类案进行特征提取与模式识别。例如，“小包公”系统采用深度学习方法，将案件划分为“定罪基准”“从重情节”“从轻情节”等模块，结合最高人民法院量刑指导意见与地方实施细则，构建理论量刑预测模型。同时，通过分析历史判决数据，系统生成实证量刑分析模型，实现“理论预测”与“实际判例”双系统交叉验证。这一技术路径有效解决了量刑标准化与地域化差异的平衡问题。

1.2.3 量刑预测与偏离度测算

在具体应用中，系统通过输入案件要素（如罪名、情节、地域）自动生成量刑建议，并基于大数据比对测算量刑偏离度。例如，安徽省怀宁县检察院使用“小包公”系统时，检察官仅需选择区域、罪名及量刑情节，系统即可在 3 秒内生成包含刑期范围、法律依据及类案对比的分析报告，并通过可视化图表展示量刑结果的合理性。此外，系统还可识别异常判决，为审判监督提供数据支持。

2 生成式人工智能辅助量刑的司法现状

本节将对生成式人工智能辅助量刑的国外应用情况和我国关于“人工智能＋司法”的现状及发展趋势展开讨论。

2.1 生成式人工智能辅助量刑的国外样态

在具体的量刑司法实践中，国外很早就开始了这方面的研究。以英美法系的美国为代表，美国司法界在人

工智能量刑辅助领域很早就展开了研究，并且也形成了较为成熟的实践体系，其在量刑方面的应用核心思路是经由算法模型来提升法官在量刑环节的规范性和效率。在适用范围上，据统计美国有超过半数的州法院都鼓励法官采用生成式人工智能辅助量刑的工具，例如宾夕法尼亚州就采用了名为“量刑风险评估工具”（Sentencing Risk Assessment Instrument）的生成式人工智能量刑辅助工具，其可以通过被告的犯罪历史社会关系等因素来对被告进行量化评分，并生成量刑建议。

2.2 生成式人工智能辅助量刑的中国实践

司法人工智能的应用是司法实现现代化必由之路。在当下开展的量刑规范化改革是人工智能辅助量刑的司法背景，与此背景相契合的是，将人工智能应用于量刑的初衷也是规范量刑。我国司法系统敏锐地注意到了生成式人工智能在司法、量刑等领域的巨大能动作用，并积极开发智能化系统，借助生成式人工智能大力推进了审判、量刑等流程的再造，全面提升了司法工作的效率，为提高实质化审判提供了有力的技术保障。据统计，全国已有 3190 家法院同时具备法条及类案推送三项功能。下文将对我国具有代表性的生成式人工智能辅助司法系统进行介绍。

北京“睿法官”系统：该系统依托司法大数据构建了“宏观-中观-微观”三级分析体系，宏观层面的构建能实现案件趋势的可视化分析，中观层面的设计能构建类型化案件裁判模型，微观层面的系统则提供个案裁判偏离度的预警。经由三重构建的系统通过将人工智能技术与司法经验深度融合，显著提升了案件办理的质效指标，有效缓解了司法裁判中的“同案不同判”现象，在规范审判权运行、统一司法裁判标准、减轻法官事务性工作负担等方面展现出显著的实践价值。

总体来说，我国目前的生成式人工智能量刑辅助功能正在蓬勃发展之中，并且在实际的运用中为司法裁判者提供了有力的帮助。

3 生成式人工智能辅助量刑的司法应用风险

本节将围绕 Chat GPT 这一代表性人工智能展开对生成式人工智能辅助量刑存在的司法应用风险的具体分析。

3.1 从技术层面审视 Chat GPT

Chat GPT 等生成式人工智能模型在训练过程中，可能受到训练数据偏见的影响，导致算法歧视。因此，在量刑辅助中，算法歧视可能导致量刑建议的不公正。人的意识是由客观存在的物质决定的，人接触事物的有限性决定了其意识的有限性，故意识不能完全呈现客观世界的面貌，其根本属性是主观的，带有个人偏好的。故法官判案的过程本质是其在带有个人偏好的意识支配下结合法律规定发挥主观能动性的过程，而这样的本质也必将贯穿于法官对 Chat GPT 的使用当中。故 Chat GPT 在量刑辅助的过程当中，其所接收到的数据来源于法官外部的输入，决定了其用于训练的数据本身就不具有完全的客观性，故 Chat GPT 在习得这种偏见后，逐步形成一套自身的逻辑推演模式，每每遇到类似的案件，Chat GPT 都极有可能对该类案件性质作出“偏见性评价”，无法达到客观中立，不仅与人们使用生成式人工智能辅助量刑以追求更加客观中立之判决的初衷相违背，还会影响司法的公正性，权威性。

除了算法风险，Chat GPT 在辅助量刑的过程中也存在一定程度的司法数据风险。首先是数据隐私泄露风险。Chat GPT 模型在训练和应用过程中，需要处理大量的司法案例和个人信息。这些数据涉及当事人的隐私和敏感信息，如果处理不当，可能导致数据隐私泄露风险，侵犯当事人的合法权益。

3.2 从道德层面审视 Chat GPT

Chat GPT 算法的不透明性使得人们难以对 Chat GPT 辅助量刑的结果进行验证和审查，当 Chat GPT 辅助量刑出现错误或偏差时，难以确定责任归属。这种责任归属的模糊性可能导致司法责任的逃避和推诿。不仅如此，Chat GPT 辅助量刑的高效性和便捷性可能导致司法人员对其产生过度依赖，从而削弱司法人员在量刑过程中的主体性和自主性。同时，量刑过程不仅涉及对案件事实的认定和法律规范的适用，还涉及对案件伦理价值的判断。Chat GPT 作为一种算法模型，难以完全理解和体现人类的伦理价值和道德观念，这可能导致算法逻辑与人类价值判断之间的冲突。

4 生成式人工智能司法风险的规制路径

本节将客观审视人工智能辅助量刑背后的法律风险，以技术与法律相结合的方式来革新我国的司法制度。

4.1 明确生成式人工智能辅助量刑的合理使用范围，坚持辅助审判原则

在法律层面，需通过立法明确人工智能辅助量刑的“工具属性”。例如，最高人民法院和最高人民检察院可出台司法解释，细化“辅助”的具体适用情形和范围。同时，为防止司法裁判人员过度依赖算法数据，可建立“双轨制”审查机制：一方面，要求法官在判决书中说明是否采纳人工智能量刑建议的情况进行说明，另一方

面，通过定期培训提升法官处理数据和算法的能力，使其能够理解量刑辅助系统的算法逻辑以及有效识别数据偏差。

4.2 确立生成式人工智能辅助量刑的具体路径，建构具体范围与限度

制定统一的司法数据标准，标准要求涵盖案件基本信息、证据材料、裁判文书等要素，确保数据格式的一致性和可交互性。建立裁判文书结构化数据元目录，对案件事实、法律依据、量刑情节等进行标准化标注。按照数据敏感程度对数据进行分类，包含公开、内部、机密三个等级，并对不同的数据设置相应的访问权限。除此以外，应当将社会主义核心价值观转化为可量化的算法指标。例如，在量刑模型中设置 “社会危害性系数” “修复性司法指数” 等参数，将 “天理、国法、人情” 作为统一整体有机统纳入算法考量参数。

4.3 建立人工智能辅助量刑的问责制度

4.3.1 多元协同监管框架

搭建“司法主体引领、技术人才支撑、社会大众参与”的协同监管制度。由司法机关来起草相关监管制度、进行合规性审核；由相关技术人才来做技术支持和风险评估工作；社会大众可以通过公示意见征询、听证等途径参与监管，还可以搭建监管协作平台，明确由司法、网信、公安等部门联合协调监管。

4.3.2 责任追究机制

首先，明确人工智能辅助量刑过程中各方的责任承担问题。开发者应当对算法的缺陷、数据错误等技术问题承担责任；司法机关对系统使用不当、监管缺失等监管问题负责；法官对最终裁判结果承担法律责任。建立 “责任倒查” 机制，对于因人工智能错误导致的裁判偏差，依法追究相关机关主体的责任。

参考文献

[1]朱鹏冀.高频多尺度编码的迭代双目立体匹配[D].电子科技大学,2024.
[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(06):98-117.
[3]张梓溪.人工智能辅助量刑的困境及突破路径[D].广东海洋大学,2023.
[4]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,(03):60-70.
[5]胡馨匀.人工智能辅助量刑的法律问题研究[D].西南政法大学,2021.
[6]李小敏.ChatGPT 在教学中的应用前景及风险防范[J].现代商贸工业,2025,(03):252-254.
[7]郭碧娟,热依拉·斯迪克.人工智能辅助量刑的风险与规制[J].南方论刊,2024,(12):80-81+109.
[8]莫皓.AI 辅助量刑制度建构的困局与纾解之道[J].西南民族大学学报(人文社会科学版),2024,45(02):55-63.
[9]钱洪伟,王旭,高宁.生成式人工智能 ChatGPT 风险形成机理与防范策略研究[J].中国应急管理科学,2024,(11):112-122.
[10]杨建武,罗飞燕.类 ChatGPT 生成式人工智能的运行机制、法律风险与规制路径[J].行政与法,2024,(04):101-115.
[11]张武军,王嘉铎.人机共融到人机共荣——以 ChatGPT 为例论生成式人工智能生成物的可著作权性问题[J].电子知识产权,2024,(01):35-43.
[12]张熙,杨小汕,徐常胜.ChatGPT 及生成式人工智能现状及未来发展方向[J].中国科学基金,2023,37(05):743-750.
[13]彭海青,于坤.人工智能辅助量刑建议的缺陷审思[J].数据法学,2023,4(01):157-174.
[14]左卫民.AI 法官的时代会到来吗——基于中外司法人工智能的对比与展望[J].政法论坛,2021,39(05):3-13.
[15]卞建林.人工智能时代我国刑事诉讼制度的机遇与挑战[J].江淮论坛,2020,(04):150-155+193.
[16]马长山.司法人工智能的重塑效应及其限度[J].法学研究,2020,42(04):23-40.

作者简介：龙颖（2000 年-），女，汉族，四川南充人，西南科技大学法学院，研究生在读，研究方向：中国刑法等
胡兰之（2001 年-），女，汉族，四川达州人，西南科技大学法学院，研究生在读，研究方向：中国刑法等
王笳薪（2001 年-），男，汉族，四川绵阳人，西南科技大学法学院，研究生在读，研究方向：中国刑法等
胥兆伟（2000 年-），男，汉，四川绵阳人，西南科技大学法学院，硕士研究生，研究方向：中国刑事法

基金项目：西南科技大学研究生创新基金资助（24ycx1170）Supported by Postgraduate Innovation Fund Project by Southwest University of Science and Technology（24ycx1170）