Open Journal of Legal Science 法学, 2026, 14(5), 222-231
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2026.145154

Hans汉斯

人工智能司法应用中的算法决策风险及其本土防范研究

宋晨菲

山东大学法学院，山东 青岛

收稿日期：2026年4月16日；录用日期：2026年4月30日；发布日期：2026年5月28日

摘 要

在司法数字化转型的浪潮中，人工智能算法正深度融入司法实践的核心环节。其在裁判辅助、风险评估、流程优化等领域的规模化应用，不仅显著提升了审判效率，更通过标准化推理模型推动裁判结果趋向统一，为智慧法院建设注入了技术动能。然而，技术赋能的背后潜藏着多重风险，对司法公正构成新的挑战。控辩双方在算法运用能力上的悬殊差距，可能引发诉讼资源分配失衡，导致程序正义的形式平等受损；司法数据的分散化存储与非标准化采集，直接造成算法模型训练数据质量不足，使得偏差性裁判时有发生；算法“黑箱”的不可解释性，既削弱了司法裁判的说服力，也为责任追溯设置了技术障碍；而过度依赖自动化处理，则可能稀释司法程序的人文价值，动摇程序正义的核心地位。基于本土司法实践需求和我国司法实践的特殊性需求，亟需构建系统化的风险防范体系。研究提出强化算法辅助定位、优化司法数据审查与采集、建立分层责任制度及推进法律价值数据化的防范体系，这一防范体系的构建，旨在为平衡技术赋能与正义维护的矛盾提供理论支撑，为智慧法院的可持续发展提供实践范本，以此为智慧法院建立过程中的技术赋能同维护正义之间矛盾解决工作给予理论支撑和参考范本。

关键词

人工智能，司法算法决策，风险识别，本土防范

Research on Algorithmic Decision-Making Risks in the Judicial Application of Artificial Intelligence and Its Local Prevention

Chenfei Song

School of Law, Shandong University, Qingdao Shandong

Received: April 16, 2026; accepted: April 30, 2026; published: May 28, 2026

文章引用: 宋晨菲. 人工智能司法应用中的算法决策风险及其本土防范研究[J]. 法学, 2026, 14(5): 222-231.
DOI: 10.12677/ojls.2026.145154

Abstract

Amid the wave of digital transformation in the judicial field, artificial intelligence algorithms are deeply integrated into the core links of judicial practice. Their large-scale application in areas such as judicial adjudication assistance, risk assessment and process optimization has not only greatly improved trial efficiency, but also advanced the unification of adjudication results through standardized reasoning models, injecting technological momentum into the development of smart courts. Nevertheless, multiple hidden risks lie behind technological empowerment, posing new challenges to judicial justice. The huge gap between the prosecution and the defense in the capacity of algorithm application may lead to the imbalance in the distribution of litigation resources and undermine the formal equality of procedural justice. The decentralized storage and non-standard collection of judicial data directly result in poor quality of training data for algorithm models, giving rise to biased adjudication outcomes on a regular basis. The inexplicability of the algorithmic "black box" weakens the persuasiveness of judicial judgments and creates technical obstacles for accountability. Furthermore, excessive reliance on automated processing may dilute the humanistic value of judicial procedures and shake the core foundation of procedural justice. In response to the demands of localized judicial practice and the unique characteristics of China's judicial system, it is urgent to establish a systematic risk prevention framework. This study proposes a prevention system covering clarifying the positioning of algorithm-assisted adjudication, optimizing the review and collection of judicial data, establishing a tiered accountability mechanism, and promoting the dataization of legal values. The construction of this system aims to provide theoretical support for balancing technological empowerment and justice safeguarding, offer practical models for the sustainable development of smart courts, and furnish theoretical references and practical examples for resolving conflicts between technological application and justice protection in the construction of smart courts.

Keywords

Artificial Intelligence, Judicial Algorithmic Decision-Making, Risk Identification, Local Prevention

1. 问题的提出

随着人工智能技术的进步，某些曾被认为仅有人类能胜任的任务，现在正逐渐通过智能机器得到辅助乃至取代，这一现象持续强化着人们对于人工智能能力的信任与认同。同样，在法律界，人工智能技术的进步激发了人们对于司法人工智能的更大期望，推动了利用人工智能辅助司法判断的实践。在司法数字化转型和人工智能国家战略不断推进的大环境下，算法驱动的决策模式逐渐成为司法实践的重要支撑，最高人民法院出台的《关于规范和加强人工智能司法应用的意见》给算法辅助给予了制度上的支持。安徽省检察机关推行的“智慧量刑系统”，依靠大数据分析来生成精确的量刑建议，多地法院采用智能分案系统来改善“简案快办”的流程，这些做法明显提升了司法效率并且增强了裁决的一致性，算法技术越加深入地介入到司法决策体系当中，它所蕴含的风险也逐渐显现出来，控诉方凭借大数据分析和证据挖掘所形成的算法技术优势，有可能会打破传统的控辩平衡，加大刑事诉讼中的结构性失衡现象，司法数据表现出碎片化、非结构化以及公开受限等特征，使得算法模型很难做到准确适配，进而引发决策

偏差，算法“黑箱”性质削减了法官的主观判断能力，审判权存在被技术取代的危险，自动化流程过分依靠数据处理指标，这或许会致使程序正义和庭审质量下降，从而损害司法权威和公信力。

我国智慧法院创建已有诸多成果，但是《新一代人工智能发展规划》[1]中提倡的追溯与问责制度，并未在司法系统中得到全方位落实，《中华人民共和国数据安全法》和《中华人民共和国个人信息保护法》虽给予了司法数据治理相应的法律保证，却仍旧缺少全面性的政策支撑，面对此种情形，在提升司法效率的同时有效防止算法决策产生风险，如何创建出合适的、契合本土化司法需求的风险防范体系也已成为数字法治建设进程中的一大关键难点问题。本研究立足诉讼结构理论与司法数据治理实践，系统剖析算法决策的风险生成机制，尝试从制度层面提出兼顾技术理性与司法伦理的本土化解决方案，为人工智能司法应用的规范发展提供理论与实践参照。

2. 人工智能算法决策的司法应用场景及价值

2.1. 应用场景

(一) 裁判辅助

1) 量刑建议

人工智能依靠大量历史案例数据，给司法实践给予类案裁判参照，明显改善量刑建议的科学性和精确度，通过形成机器学习模型来剖析司法决策机制，产出准确的量刑建议计划，削减“同案不同判”情况。利用智能算法做到历史案例的快速查找和搭配，给法官给予类案参照根据，改良审判进程，改善判决统一性，凭借大数据分析和算法模型，对刑事量刑要素展开量化评价，给法官给予量刑参照根据。这种技术创新既改善了司法运作效率，又通过规范化的量刑程序加强了司法公信力[1]。在实践层面，人工智能辅助量刑系统在不少地方的法院和检察院实现了逐步推行。以安徽省检察机关自主研发的“智慧量刑平台”为例，此系统凭借大数据技术对案件数据展开深入挖掘和综合分析，给司法裁量给予科学依据，既加快了办案速度，又提高了判决结果的准确性和公正性[2]。“智能辅助量刑裁决系统”依靠大数据分析技术，可以快速处理大量案件数据，并创建科学的量刑建议模型。该系统在法官输入案件信息之后，利用相似案例匹配算法自动形成量刑参考区间，从而解决由于地域差别或者经济发展水平不均衡引发的量刑不平衡状况。

2) 文书辅助

算法系统把自然语言处理和深度学习技术融合起来，要达成案件事实的自动提取，并且形成起诉书，辩护词这些法律文书，极大改进法律服务的效率和精确度。它的关键技术是对海量法律文献展开深度学习和模式识别，系统凭借剖析历史判例数据，拆解法律文书的规范结构和常用表达形式，准确提炼案件关键要素并搭建逻辑框架。在生成起诉书的时候，用户要输入案情要素，比如事实陈述，法条引用，诉讼请求等等。系统依靠事先设定好的模板以及有关法律法规，就能自动产生符合格式要求的初稿文档，而且会联系类似案例给出专业意见或者参考看法。该自动化处理模式极大地缩减了律师手工撰写法律文书的时间，其文档生成速度相比传统方法大约提升了 30%到 50%，从技术角度来说，算法系统依靠自然语言处理技术来达成法律文书的智能化生成，基于 BERT 等预训练模型创建的双向 Transformer 架构，可以精确把握法律文本里的语义联系，在法律推理以及文书自动形成方面有着明显的优势[3]。算法依靠机器学习技术来深入分析法律文本，可以准确找出其中的漏洞或者格式上的小问题，还能自动完成修复工作，保证文档符合规范。在商业合同产生争议的案子里，用户只要给出主要要素，就能很快得到结构合理、思路清晰的诉状文本，这样就大大缩减了文书的制作时间。这项技术改善了文书撰写时的准确性，

[1]https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

经过很多次改进之后，它生成文书的能力已经慢慢接近甚至超过了人工撰写的水准。

(二) 风险预测

1) 再犯风险评估

依靠犯罪主体背景数据的分析框架，所设计的算法可以评判一个人再次犯罪的可能性，并且能够给预防干预措施的制订给予科学的依照。这项技术的关键之处在于把大数据处理同机器学习模型结合起来，通过整合多种不同来源的数据，来深入探究隐藏的联系模式，进而形成起比较准确的风险预估模型[4]。算法依靠大规模犯罪数据库展开训练，细致探究特定年龄层次，家庭环境以及犯罪类型与再次犯罪之间的联系。通过研究得知，青少年囚犯以及失业人员存在较高的再次犯罪概率。此项技术不但明显改善了犯罪预测的准确性，而且给司法决策给予了数据支撑的科学依据。从技术架构角度来讲算法结合了统计学分析手段和机器学习模型，另一项有关的研究把犯罪的时空分布特点同朴素贝叶斯分类器联系起来创建预测模型，它的准确率相比传统方法有了很大的改进[5]。预测性警务通过网络分析犯罪者的社会关系，识别潜在风险个体，并结合实时监控调整警力部署。这些技术手段使执法部门能够更精准地分配资源，预防犯罪热点区域的再发。

2) 案件结果预测

搭建起历史案例数据作为基础的判决预测模型，为司法实践给予智能化辅助，这成为法律人工智能领域的重要研究方向。这种方法模仿法官在实际审判时参照过去判例的行为模式，融合机器学习技术来设计预测框架，给审判决策给予数据支撑。它的基本思路包含这些环节：从大量历史案例数据库中挑选出和待决案件最相似的 Top-k 样本，利用向量化技术提炼关键特征信息，凭借案例相似性评估算法改善特征表达形式，然后创建起多任务学习模型，把相关变量综合起来，从而提高预测准确度。有关研究显示，Han Zhang 等学者所提出的集成框架可以精准地描绘案例之间的内在联系，而且通过多层神经网络达成对判决结果的高效联合预测，在多个公开法律数据集上的表现远胜过传统的单个模型方法[6]。此外，模型还需考虑案件类型、法官历史倾向、司法管辖区、法律论点等关键变量，例如通过剖析法官以往的投票记录或者裁决偏好来改善预测精准度，这样的改良既有益于加强司法决策的公正性，又可以凭借数据驱动的方式帮法官在复杂情形下做到利益平衡。

(三) 程序优化

1) 自动化文书生成与案件分流

人工智能技术在司法领域应用主要集中在文书自动化生成与案件智能分流两个核心环节上，关键在于利用算法来达成案件自动分类、精准匹配法律条文以及高效产出标准化文书的目标，以此大幅度提升司法运行效率并改善资源分配情况，当下该技术已在国际范围内得到广泛使用，在刑事审判领域同样表现出明显的优势特征，在案件分流方面，依靠自然语言处理和机器学习模型，可以对案件实施智能化识别和准确归类，有效把复杂案件同简单案件区分开来。以安徽省检察机关为例，他们依靠机器学习算法对危险驾驶案件执行自动化分级管理，把简易案件和复杂案件分别归入不同的审理程序当中，这种机制既大大减轻了办案人员的工作压力，又促使“简案快办、繁案精审”的司法资源配置优化目标慢慢得以达成[7]。在文书生成环节，算法技术通过预设的法律文书模板和规则引擎，实现案件材料的自动填充与格式化。以安徽省检察机关为例，其智能语音与人工智能联合实验室开发的系统，通过提取案件证据要素并关联制式文书格式，一键生成审查报告，极大缩短了文书制作时间[7]。这种技术不仅降低了人为错误率，还通过标准化流程确保文书质量。

2) 智能案件分配

智能案件分派系统依靠自然语言处理技术，用多层架构做深入的文本信息解析，从而做到案件的准确分类和快速流转，进而改善司法资源的调配效果，它的核心技术模块包含依靠深度学习的多标签分类

算法，情感分析部分以及任务调度策略的协同工作，系统前端收到案件资料以后，利用深度学习模型找出纠纷类型和核心诉求特征，再借助情感分析工具判定案件的紧急程度，给后面的分派赋予优先级排序的依照。在特征工程阶段，平台把用户行为轨迹，律师的专业范围等多源数据融合起来，在离线和在线环境下做好数据预处理，创建案件和律师相配的推荐体系，自然语言处理技术可以分解起诉状，判决书这些非结构化的文档里的关键要素，而且推送有关案例帮助法官确定争议焦点。以上案例表明，自然语言处理 NLP 技术极大地提升了案件分配的智能化水平，而且通过削减人为干预有效地遏制了人情案以及关系案的出现，从宏观层面来讲，数字技术的应用改善了司法资源的调配效果，化解了“案多人少”的矛盾，促使司法服务朝着细致化方向发展，削减了诉讼成本，进一步维护了司法的公正与权威[8]。

2.2. 核心价值

司法实践中，人工智能同自动化技术的深度结合正在深刻改变司法运作机制，依靠算法推动的智能化手段，司法体系在效率改善和裁判一致化方面有着明显的优势。从效率改良角度来说，自动化技术通过案件流程管理，语音识别以及文书自动生成这些功能模块，大幅度缩减了审判周期并且削减了人力投入，智能审判平台达成全流程的数字化运作，包含自动立案，庭前预备，庭审辅助等等环节，有效地取代了传统的手工操作，大数据分析工具可以准确地找出关键数据，给法官给予科学的依照，从而极大地改进审判工作的效率。

就司法公正而言，算法技术依靠统一法律术语诠释以及裁判准则来削减法官个人要素对判决成果的影响，凭借标准化法律数据库，这种技术保证同类案件在不同地区和法官之间取得同样的裁决结论。智能辅助决策体系联合过往案例剖析，给法官给予客观参照根据，削减因为个人经历差别而产生的判案误差，算法透明且可追踪性强化了司法公信力，通过审计机制保持决策过程的公正性与科学性，这种一致性并不仅仅表现在量刑幅度上，而且牵涉到审查证据，规范流程等很多方面，拿运用语音识别技术当即把庭审谈话转录成文字记录来说，就能保证信息采集的精确性与完备性，技术赋予的司法统一化操作，既加强了司法权威，又给当事人赋予更为稳固可靠的法律预期。

算法在司法领域有明显应用前景，不过它会对法官实务能力产生不良影响，算法设计若出现偏差或者不够透明，就容易造成歧视性判决问题，亟需通过审查和反馈来加以约束。要想明晰算法辅助决策的法律地位，并且合理划分责任界限，就要避免因为技术瑕疵而引发的司法纠纷。从国际趋势看，欧盟《人工智能法案》已将司法辅助系统列为高风险应用，要求满足透明度与人为监督的强制性标准[9]。技术领域同样有所推进，Ribeiro 等人提出的 LIME 模型通过局部近似方法揭示算法决策的关键影响因子[10]，Wachter 等人则倡导以“反事实解释”弥补黑箱缺陷，即向当事人说明“何种条件改变将导致不同结果”[11]。这些技术方案为司法算法的可解释性提供了工具基础，但如何将其从技术标准转化为司法审查的法律规范，仍是制度构建中待解的难题。在推动人工智能技术进入司法实践进程中，要在效率提升与公平保障，技术创新与人文关怀之间找到恰当的平衡点，依靠健全法律法规体系和技术治理体系，让其发挥出全部潜力。

3. 司法算法决策的风险识别

3.1. 司法结构存在失衡

人工智能算法在司法领域应用存在诉讼结构失衡风险，大多数国家采用以法官为中心，原告和被告为对立面的传统三元诉讼模式，目的是保障控辩双方平等对抗，促进案件事实全面审查，维护法官中立性与公正裁判，此时各参与方权利义务分配较为平衡，但技术发展后这种模式的弊端逐渐显现[12]。工智能技术被深度运用，控诉方在资源获取上取得明显优势，依靠大数据分析、智能证据挖掘这些工具，控

诉方可以迅速搭建起关键证据链，但辩方因为技术障碍或者资源短缺很难做出有力回应，这种“技术沟壑”极大降低了辩方的辩护效率，使得诉讼流程慢慢偏离传统意义上的“平等对抗”，变成一种以技术为主要特征的不对称博弈形式[13]控方的电子化、智能化发展可能打破刑事诉讼中“控辩平衡”的基础，导致审判权与检察权过度强势，甚至出现“谁强谁有理”的实质不公。

人工智能技术的引入可能对法院的独立审判权形成一定冲击。一方面，在法院审判过程中，人工智能的技术特征造成法院审判权中法官的主观权威地位受到挑战；另一方面，在司法裁判中对信息量与处理速度的过度追求，可能造成司法裁判中的信息过载、司法决策权力让渡给算法决策系统、司法机关过度依赖技术公司等一系列问题[14]。人工智能所具备的高效性可能会使得法官过于依赖技术工具，从而影响到法官在司法审判过程中充分发挥主观能动性，对整个案件的事实进行全面审查，当人工智能被设置成“量刑辅助系统”，法官就容易信任人工智能的权威性而放弃自己的独立判断，这将对司法的公正与权威性造成影响，同时也要警惕“功能外溢”的风险：如果人工智能技术能够代替法官行使职责，不仅有损于司法独立性，而且也会导致公众失去对司法系统的信任。司法体系里存在的结构性瑕疵常常会减弱当事人权益保障的效果，而以人工智能算法为核心决策机制则大概会进一步缩减当事人的诉讼参与范围，在自动化审判系统当中，当事人很可能会被当作“数据输入”工具看待，当事人作为独立诉讼主体的基本地位以及陈述权、辩论权等程序性权利或许会被算法流程的主导作用所削减，这种无视程序正义的做法会使司法实践变成技术导向的工具，而不是公民权利得以保障的平台。

3.2. 司法数据来源准确性低

人工智能技术在司法实践当中碰到不少阻碍，数据品质的不足成为关键限制因素。相关研究显示，我国司法数据存在碎片化，分散化以及标准化程度低等特征，这些瑕疵严重影响了算法模型的训练效率和决策精确度，司法系统内部产生的大量非结构化数据缺少统一标准，这就造成它们很难做到高效整合并深入剖析，有些法律文书没有及时公开或者公开范围有限，客观性和权威性就遭到质疑，裁判根据和推理过程的一致性问题同样使得数据的真实性与可靠性下降[10]。这种数据质量的缺陷不仅限制了人工智能对司法实践的深度理解，还可能引发算法偏见和决策偏差。从技术深层看，数据偏差导致的“偏见输入–偏见输出”现象在缺乏公平性度量机制的环境下会被无限放大。

而引入算法公平性技术理论审视，司法数据中的历史偏见会通过机器学习被固化为系统性歧视。例如，在再犯风险评估中，若训练数据包含历史上针对特定群体的严苛执法记录，算法极易习得并强化“高犯罪风险”的统计相关性，而非真实的因果关系，从而陷入自证预言的恶性循环[15]。这种歧视往往具有隐蔽性，传统的司法审查手段难以察觉。由于缺乏对数据特征的公平性约束，如未融入“人口统计均等”或“均等化几率”等公平性指标，算法模型在针对敏感属性时极易引发差别影响。现行法律规范仅关注数据的形式合法性，却未触及数据的实质公平性，导致不准确的数据在源头便埋下了侵蚀司法公正的隐患。

当下司法数据领域核心问题体现着主观性显著与结构化程度低两个方面，法官对相似案子处理时习惯使用区别化判别方法，裁判文书呈现着强烈的个人倾向色彩，与此同时法律用语模糊不清、地方特点各异且文本表述统一性不够等诸多状况并存，人工智能提取、解析司法数据的技术难度由此提升了不少，不少案件受地域条件或者法官独特风格等干扰之后推理过程表现出很大不同，这种类型的差异可能会使算法训练偏离正轨，最终结果也不合乎实际司法需要[16]。不准确的数据可能导致算法输出错误的法律结论，在再犯风险评估中，若算法仅依赖片面的数据特征，可能忽视关键变量，从而引发误判数据偏差会加剧算法偏见。这些问题不仅威胁司法独立性，还可能引发公众对人工智能司法系统的信任危机。

3.3. 司法责任不明晰

人工智能算法在司法领域使用之后，虽然提升了效率，而且改善了精准度，不过算法作出决策的时

候不透明的特性，责任归属模糊的问题成了要解决的关键。从技术深层看，当前算法的“黑箱”不仅表现为透明度不足，更在于其无法提供契合司法论证逻辑的因果解释。判决的可接受性根植于因果论证而非统计相关性，算法仅输出结论却无法回应“相似情形为何不同处理”的追问，当事人的质证权利和司法的论证义务便在实践中双双落空。从根本上说，这主要因为司法过程中的人工智能系统责任主体存在多种复杂性，算法输出的结果会受算法设计，数据输入，用户交互等多种因素影响，当出现偏差或者争议状况时，责任往往会分散到开发方，使用者，数据提供者以及监管机构等多个主体身上。由于算法的运行机制很大程度上依靠复杂的资料处理和机器学习模型，内部运作过程很难被不懂内情的人清楚地掌握，于是司法程序缺少透明度和可说明性。

《新一代人工智能发展规划》提出建立人工智能产品可追溯的责任制度，要求明确人工智能法律责任承担主体以及相关权利、义务和责任等[17]。当下，人工智能算法在司法实际应用中不断深入，它同司法责任体系的内在联系也变得越发明显起来，如果没有系统的法律规制框架，就会影响到当前司法责任制的执行，特别在依靠人工智能算法支撑的司法决策流程当中。责任主体判定，过错归责机制设计以及利益相关方权益维护等关键问题急需用制度化的形式来加以明确，责任主体界定不清不仅会干扰司法权的正常运作，而且还会降低对当事人合法权益的守护强度，进而妨碍相关行为的规范化管理和高效监督。确定了责任主体之后，人工智能算法引发的司法责任分配还要继续细化其边界，要重点研究算法决策流程里的责任划分问题，而且要形成起科学合理的分担机制，这样做既有益于维持司法体系的公平性和权威性，又能够切实保障相关利益主体的合法权益。

3.4. 司法程序价值缺失

虽然人工智能算法在司法实践领域具有提高效率、保证准确性的潜力，但是人工智能算法也存在相应的风险。透明性、民主性、中立性、权威性作为维护司法公正的重要因素，正在面临前所未有的挑战。当以算法为基础的“自动化决策系统”代替传统司法程序时，程序的严肃性、对抗性被削弱，法庭环境的庄重性也被打破。虽然在线诉讼给当事人参与诉讼带来方便，但是很难再现实体庭审的“剧场效应”，从而动摇公众对司法权威的信任基础[18]。程序的亲历性、规范性与严密性亦因技术简化而受损，如异步审理可能违背直接言词原则，线上作证的有效性与证据展示的质证效果存疑；而算法系统仅输出结论、不展示推理过程的技术特性，更使当事人无从追问“何种情节改变会导致不同结果”，申辩权与异议权难以有效行使。这种“去程序化”倾向，实质上是对司法程序价值的系统性侵蚀。

当下，人工智能算法在司法实践中被应用时碰上程序性规范缺少的瓶颈，人工智能算法算是一种也许会对公民基本权利造成重大影响的高风险技术工具，实际操作当中缺少系统的程序性约束，司法活动表现出一定的随意性，特别在适用条件，运行机制以及操作流程等层面，有关规则和标准尚不完善，不能给外部主体给予清晰的信息指引，这种状况既妨碍了人工智能算法决策进程的规范化发展，又也许会引发潜在的法律风险和伦理争端。司法主体如果在缺乏实质审查的背景下，将人工智能算法所产生的决策结果当作终局裁决参考，这样可能会伤害到当事人的权益，而且会对司法公信力形成负面影响，在程序正义这一视角之下，人工智能技术的“工具化”现象冲淡了司法程序的真正意义，作为法律施行的重点步骤，司法程序既是保证法治秩序的关键手段，也是达成社会公平正义的主要途径，当算法决策占据主导时，程序对抗性与规范性的核心特质会被缩减成单纯的数据处理速度评定标准，进而致使当事人在心理上产生对程序公正性认同度及信心不足。

4. 风险防范的本土制度构建

4.1. 强化算法决策在司法运用中的辅助性定位

人工智能具有黑箱决策的性质，将司法决策权转移到代码或程序员的手中，人类的司法决策权将被

剥夺或减少，司法审判中的问责制将被削弱。新型重大疑难复杂案件中，应确保人的绝对控制地位，避免技术主导风险，智慧法院建设中，保障司法权的主导地位至关重要，特别是在处理新型重大疑难复杂案件时，需要法官在决策和判断中拥有主导权，而智能化系统仅仅作为辅助工具[19]。在司法实践方面，准确把握算法决策的辅助功能属性，是形成本土化风险防范机制的关键部分，这种属性既是人工智能技术融入司法系统的基本要求，也是保证司法公平正义与效率的重要根基，《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）[2]中规定，应突出人工智能技术的辅助特征，一方面要明晰其在司法流程中的职能范围，另一方面也要捍卫法官在审判过程中的自主裁量权利，把人工智能算法纳入司法体系的主要意图在于依靠数据支持与智能化手段优化裁判品质，并不是要取代人类主体地位从而达成自动化决策。

通过立法或者政策性文件来确定人工智能在司法领域里所扮演的辅助职能定位以及功能边界，这有助于防范技术应用可能会给司法独立性带来潜在的风险，要突出强调 AI 系统只是辅助工具，并不是法官裁判的主体，法官依然要对裁判结果负责，应该形成人机协同决策机制，依靠可解释算法和分层透明规则设计，使得法官可以理解 AI 建议背后的逻辑依据，做到技术赋能与司法公正之间动态平衡。具体而言，可要求司法 AI 系统在涉及当事人重大权益时提供“反事实解释”，即展示若某一关键情节发生变化，决策结论将如何不同。例如，系统拒绝取保候审时，不应仅给出风险分数，而应说明何种条件下风险值会降至阈值以下。这种解释范式不仅破解了“黑箱”，更高度契合司法实践中“区别技术”的思维逻辑，构成法官心证的有力外化。立法上，应将反事实解释能力作为智慧司法系统通过验收和审计的法定技术要件，将技术上的可解释性转化为法律上的可论证性，使算法建议真正纳入裁判说理的逻辑框架[20]。就司法实践当中人机协作模式所遭遇的适配性及灵活性欠缺状况而言，亟需改良有关机制的设计，可以考虑给各级司法机构赋予在特定案由或者层级范围之内自行调整人机决策权重的权力，在审理疑难复杂的案件期间，加强法官的主体裁量作用，而当处理一般的程序性案件的时候，突出人工智能的技术辅助效能[21]。人工智能算法帮助法官裁决案件的时候，要给它修正或者否决算法建议的权利，当人工智能系统给出的意见和法官的主观判断有冲突的时候，一定要按照“以人为主、科技为辅”的原则，优先保证人工决策的权威性和主导地位[22]。法官仍然可以坚持自己的意见，基于自由裁量权作出决策，事后再根据情况由审判监督部门进行审查。

4.2. 优化司法数据信息采集与审查

在司法数据采集方面，必须严格遵循合法性与规范性原则，依据《中华人民共和国数据安全法》和《中华人民共和国个人信息保护法》以及《中华人民共和国网络安全法》等法律法规的规定，境内收集到的个人信息原则上要存放在国内，从这个角度来讲，司法数据采集工作务必全面履行有关义务，在数据采集之初就要保证数据来源的合法合规，并且执行必要的风险评估和安全审查步骤，针对涉及到隐私的语音数据等敏感信息，其采集过程更应谨慎推进并加以细致控制，整个数据处理流程之中，应当明确划分权属关系，真正保障各个步骤的操作符合法律法规要求，进而有效地防止因为权属问题而产生的法律风险[23]。数据入库流程设计要融合数据确权管理模块，保证数据来源合法，权属清楚，存储和应用的时候执行权限控制和符合性检查，审核阶段，先由司法事务承办人员和文书撰写人员自己检查，重点是法律适用偏差，理解误差之类的专业技术问题，之后把资料交给专门负责司法文书数字化的技术团队再次检查，去掉不符合规范的文件，对找出的问题立刻给出改正意见，文书正式存入司法数据库以后，要创建动态监督机制和责任追溯体系，这样就能在察觉到也许存在的错误或者违规行为的时候，立即采取改正办法并落实追责制度。

[2]https://www.court.gov.cn/zixun/xiangqing/382461.html

4.3. 建立司法责任承担制度

国家应对人工智能算法决策造成的司法失误承担系统责任，这种责任源于算法在司法实践中的核心地位，算法的运行实际上就是国家法律职能的执行过程，司法机关作为权威执法机构，不仅要对算法决策带来的法律后果负责，还要完善对算法运行机制的监管和规范措施，法官和其他司法人员也应对基于算法做出的不当裁判负责，在此情况下，建立高效的内部监督机制至关重要，这有助于确保人工智能技术在法治轨道上有序发展。

从国家治理体系的角度出发，要规避人工智能算法决策所引发的风险，就需要形成类似于“国家赔偿”的责任分摊架构，可以参照我国《国家赔偿法》里的有关条文，针对特定情况下的责任划分予以细致化规划，按照现有法律法规的规定，在司法机关执行诸如侦查，检察或者审判之类的活动时，如果侵犯了公民的合法权益，受害者就有权利依法提出申请国家补偿的要求，处在这样的环境之中，司法部门要进一步提升对涉及人工智能算法应用案件的监督与引导职责，从而保证公正执法和合法维权目的达成。一旦司法机关在对人工智能算法决策系统实施监督之时有所失误，比如没有保证系统按时接受检查，合理测试，也没有做好全面评价，或者不采取适当的安全保障手段，那么就应当为其造成的结果承担相应的责任，在利用人工智能去协助裁决案件时，若没有遵守有关程序的规则，涉及司法的那些单位就会受到程序上的缺陷所带来的法律后果。

4.4. 实现司法决策的价值判断数据化

人类法官在对司法案件的处理过程中除了有对司法理论的把握，还有个人的价值判断、司法实践经验等。当前人工智能算法虽然无法做到独立的价值评判，不过这个领域的发展状况依然被很多人所关注，最高人民法院已经明确要求各个法院加大对司法人工智能关键技术的研发投入力度，从而改进其在实际审判过程中的应用效果，在模仿人类法官作出决策的时候，人工智能系统碰上的最大难题之一就是怎样妥善处理好各种多目标权衡及优先级排序问题，日本学者平田勇人依照法律公理化理论，给出了利用价值函数公式化模型来解决这种冲突的技术方案，给改善人工智能算法在复杂法律情形下的决策逻辑给予了关键参照[24]。创建法律价值和规范的量化评价体系，可以做到司法价值和制度层级的科学排序，在刑事诉讼过程中采用人工智能来辅助决策的时候，要把公平性放在效率之前予以优先保护，这样既能塑造符合司法公正需求的人工智能算法模型，又能为多维度司法目的出现矛盾时给予恰当的选择标准。

5. 结论

本文的研究依托司法数字化转型和人工智能国家战略的宏观背景，探究人工智能算法在司法中的应用风险产生机制以及本土化的防控对策，算法决策已经渗透到裁判辅助、风险预测和流程优化等各个环节之中，诸如安徽省的“智慧量刑系统”，借助大数据技术提高量刑准确率，各地的法院通过智能分案机制来改善资源配置情况，司法效率得到明显改善，判决一致性也变得更高一些。技术赋权产生的风险也不可忽视。控辩双方由于各自拥有的技术程度存在差异，从而进一步推动诉讼中的权力失衡状况，而数据孤岛又造成模型训练出现偏差情况，“黑箱决策”会削减法官的审判独立权，“自动化操作”甚至会削弱整个程序所体现出的公正价值。

针对上述问题，研究构建了四维本土化防范体系：明确算法的司法辅助定位，建立“法官主导–算法辅助”的决策机制，避免技术取代司法主体地位；构建司法数据全生命周期治理体系，解决数据采集、存储、使用中的标准化与合规性问题；参照《国家赔偿法》逻辑，创设“国家概括责任–司法机关监管责任–法官主体责任”的分层架构，厘清决策失误的责任归属；借鉴法律价值函数理论，探索司法公正、效率等价值的量化赋值标准，实现技术理性与司法伦理的有机统一。未来，人工智能在司法领域应用要

达成技术革新与制度约束之间的动态平衡体系，只有如此，才能使它由“效率优先”向“公平正义与效率并重”转变，从而给数字法治创建赋予具有本土特色和时代意义的更新颖的解决办法。

参考文献

[1] 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(1): 157-174.

[2] 李浩. 检察事业插上“智慧翅膀” [EB/OL]. http://www.anhuinews.com/ahkj/qwfb/202303/t20230316_6730050.html, 2025-07-03.

[3] Devlin, J., Chang, M.W., Lee, K., *et al*. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

[4] 段黎宇, 张钰莹. 人工智能在侦查中的运用构建与职能定位[J]. 湖南警察学院学报, 2020, 32(6): 45-53.

[5] Gouse Basha, S.K., Ramana, A.V. and Murali Krishna, B. (2023) Predictive Analytics for Crime Prevention by Using Machine Learning. *International Journal of Research Publication and Reviews*, **4**, 59-63.

[6] Zhang, H. and Dou, Z. (2023) Case Retrieval for Legal Judgment Prediction in Legal Artificial Intelligence. *Proceedings of the* 22*nd China National Conference on Computational Linguistics*, Harbin, 3-5 August 2023, 801-812.

[7] 缪成, 汪迎兵, 李宝善. 机器学习算法在法律文书制作繁简分流中的运用——以危险驾驶案件审查报告自动生成技术为视角[J]. 人民检察, 2021(2): 66-68.

[8] 尹泽贤. 创新大数据在案管工作中的运用[J]. 人民检察, 2019(15): 77-78.

[9] Regulation 2024/1689 of the Eur. Parl. & Council of June 13, 2024 (EU Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[10] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) Why Should I Trust You?: Explaining the Predictions of Any Classifier. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD'16), San Francisco, 13 August,2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[11] Wachter, S., Mittelstadt, B. and Russell, C. (2018) Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, **31**, 841-887. https://doi.org/10.2139/ssrn.3063289

[12] 谢佑平, 万毅. 刑事诉讼法原则: 程序正义的基石[M]. 北京: 北京法律出版社, 2016: 206-208.

[13] 郑曦. 人工智能司法运用的风险与规制[N]. 人民法院报, 2021-01-21(006).

[14] 张凌寒. 智慧司法中技术依赖的隐忧及应对[J]. 社会科学文摘, 2023(6): 18-20.

[15] Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) Machine Bias: There's Software Used across the Country to Predict Future Criminals and It's Biased against Blacks. ProPublica.

[16] 黄悦晨. 人工智能在司法裁判领域应用的局限性及其改进对策[J]. 法学(汉斯), 2023, 11(3): 1806-1810.

[17] 马驰. 谁可以成为法律主体——兼谈人工智能的法律主体资格问题[J]. 甘肃社会科学, 2022(4): 129-141.

[18] 陈敏光. 司法人工智能的理论极限研究[J]. 社会科学战线, 2020(11): 194-204.

[19] 王秀平. 人工智能司法应用法律风险及法治应对[J]. 政法论丛, 2024(2): 151-160.

[20] 黄国栋. 比较法视野下智慧法院建设的中国经验、实践困境与路径优化[J]. 法律适用, 2023(3): 129-138.

[21] 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.

[22] 张玫瑰. 司法裁判中人工智能应用的限度及规制[J]. 政法论丛, 2023(5): 128-138.

[23] 韩文. 语音数据法律风险防范的本土制度构建[J]. 法商研究, 2023, 40(5): 90-102.

[24] 平田勇人. 信义原则与基础[M]. 东京: 东京成文堂, 2006: 287-289.