# 重塑信任:DeepSeek类生成式人工智能嵌入司法审判的底层逻辑和优化路径*

崔立红　李昶郴

(山东大学 法学院,山东 青岛 266237)

**摘　要**:在司法审判过程中应用DeepSeek类生成式人工智能,一方面,可以减轻法官的认知负担,避免认知过载现象,提高司法效率;另一方面,可以克服法官的认知偏见,防止偏见驱动现象,确保司法公正。但信任问题仍是DeepSeek类生成式人工智能嵌入司法审判的主要阻碍,主要表现为继承数据歧视、"黑箱"问题压缩解释空间和人工智能参与决策责任归属不明等问题。为此,要重塑司法实务中对DeepSeek类生成式人工智能的信任,可以通过明确DeepSeek类生成式人工智能的使用边界、发展可解释人工智能以及确立法官主体的决策责任三种路径,构建"人机协同"的新型审判范式。

**关键词**:DeepSeek;可解释人工智能;认知心理;决策责任;人机协同

**中图分类号**:D926;TP18　　**文献标识码**:A　　**文章编号**:1672-335X(2025)05-0083-11

**DOI**:10.16497/j.cnki.1672-335X.202505008

## 一、问题的提出

DeepSeek、ChatGPT等生成式人工智能在司法领域的应用不再是未来的设想,而是当前的现实。2022年11月,OpenAI公司推出了生成式人工智能模型ChatGPT。ChatGPT凭借其自身强大的语言理解与生成能力,在全球范围内引发了广泛关注。2023年12月,英国司法办公室发布一项声明,使法官利用大语言模型辅助工作的做法合法化。[1]这是全球首次有政府文件正式讨论司法系统中大语言模型的使用,随附的文件明确说明了法官应如何负责任地利用人工智能技术。中国科技公司迅速响应全球生成式人工智能的发展趋势,相继推出Kimi、豆包等本土化大语言模型。国产大语言模型DeepSeek更进一步创新技术,用极低的算力成本实现比肩全球一线预训练大语言模型的能力,还首次公开使用强化学习(reinforcement learning)作为推理模型的可能路径,为大语言模型从"会说"向"会想"迈进提供了新的方向。至此,国产大语言模型DeepSeek的崛起,直接标志着生成式人工智能在我国司法领域的应用已具备坚实技术基础,展示出巨大应用潜力。

"DeepSeek类生成式人工智能"是指以DeepSeek为代表的一类具备自然语言生成能力、能够在特定语境中模拟法律文本编写与语义推理的人工智能工具,其核心特征在于通过对海量语料的训练,实现语言生成、类案匹配与事实重构,在司法审判活动中逐步承担从认知辅助、程序性支持到初步审查支持的实质性功能。自2016年最高人民法院提出智慧法院建设目标以来,法院持续关注人工智能技术在司法工作中的应用,目前,已探索出全流程网上办案体系、电子卷宗机制、智能辅助办案系统及用于支持智

* **收稿日期**:2025-06-06

**基金项目**:山东省社会科学规划重点项目"贡献原则下人工智能生成物著作权归属研究"(25BFXJ02)

**作者简介**:崔立红(1968-　),女,山东青岛人,山东大学法学院教授,博士生导师,主要从事知识产权法、科技法研究。

能分析的相关司法大数据库等人工智能辅助工具。[2]然而,这类工具更倾向于承担流程管理与辅助裁判功能,其核心技术仍属于传统的规则驱动式人工智能,主要依靠预设的规则和大量的训练数据来工作。相较之下,DeepSeek 类生成式人工智能具备强大的自然语言理解与内容生成能力,能够参与更复杂的法律语义处理工作。例如,DeepSeek 强大的语言处理能力使其能够在法律文本分析、案件预测、判决草案生成等方面为司法工作提供重要支持,从而提升司法效率和准确性。因此,在功能定位上,DeepSeek 类生成式人工智能不再仅局限于进行流程管理或文书编写,而是有能力嵌入事实认定、类案比对与裁判辅助等核心环节。虽然 DeepSeek 类生成式人工智能在司法领域的应用前景广阔,但是它仍面临着诸多挑战,尤其是在可解释性、数据歧视和责任归属方面存在的问题使其在司法领域的适用面临信任危机。因此,确保人工智能工具的决策过程透明、防范人工智能模型可能带来的偏见以及明确法官与人工智能之间的责任划分,都是在司法工作中使用 DeepSeek 类生成式人工智能工具时必须重点关注的问题。

## 二、DeepSeek 类生成式人工智能嵌入司法审判的底层逻辑

DeepSeek 类生成式人工智能协助法官的主要方式之一是自动化行政任务。一方面,人工智能可以完成搜索大量法律文件和案例的繁重任务,减少法官在完成这些初步任务上花费的时间和精力,使他们能够更多地专注于实质性的法律问题和审判工作。另一方面,人工智能有减少人类认知偏见影响的潜力。法官和其他人一样,其思维过程容易受到认知偏见的影响。相较而言,经过多样化数据集训练的人工智能系统可以进行更为客观的案件分析,实现更公平的裁决。

### (一)减轻认知负担

在大陆法系中,法官在评估、解释和综合证据及证词方面扮演着积极和集中化的角色,他们是司法认知活动的核心主体,负责作出司法决策,最终形成具有法律约束力的裁判。与对抗性诉讼程序不同,在大陆法体系中,案件的程序并非仅由律师之间的对抗性论据和反论据来推动,而是要求法官独立对每一项证据进行评估,判断其证明价值。这种程序设计赋予了法官较大的权力,同时也带来了较大的认知负担,法官必须基于客观、透明且充分论证的证据作出裁决。然而,随着案件复杂度的上升,文件、法医学证据、证人证言、专家证词数量的增多及相互依赖性的增加使证据综合和无偏决策变得更具挑战性。当法官面对相互缠绕的证据网络时,其认知活动必然受到工作记忆容量、注意力分配机制与情绪调节能力等生物学因素的制约。

在认知心理学领域,人类在记忆和处理信息方面的局限性已得到广泛的研究。人类大脑的前额叶皮层被认为是执行功能的核心区域,其处理多重任务的能力受限于神经递质系统的代谢效率。[3]例如,当证据评估涉及超过 5 组独立变量间的非线性交互作用时,大脑默认模式网络与背侧注意网络之间的资源竞争将导致认知控制能力显著下降。[4]根据心理学上的经典理论米勒法则,人类在任何给定时刻只能可靠地处理约 7 ± 2 个(即 5—9 个)信息项。[5]这种认知限制在司法工作中的影响尤为显著,因为法官必须分析多个相互关联的证据,以确保作出公平和明智的决策。随着案件复杂性的增加,法官的认知负担加重,很难同时评估多个证据。当证据之间存在复杂的依赖关系时,尤其是当证人证言相互证实、证人证言发生冲突、法医报告与其他文档记录相互矛盾等情形出现时,法官在记忆和处理信息过程中会承受极大的压力,认知过载便更加严重。

认知过载往往会使法官依赖认知捷径以简化复杂的判断过程,从而易产生判断偏误。心理学研究表明,人类决策多依赖启发式(认知捷径)作出判断,以简化复杂的判断过程。[6]启发式是指人类在面对复杂决策时,通过简化问题或根据过去经验作出快速判断。这种思维方式虽然可以提高决策效率,但容易导致偏误。在案件审理过程中,法官因信息量庞大、时间紧迫或案件复杂,会过度依赖启发式决策方法,导致很多决策并非基于详尽的理性分析,而是通过自动化的、直觉性的方法作出。这种快速、自动的

判断(即快速思维),比起费力、理性、细致的思考(即慢思维)更容易被法官采纳。[7]例如,法官可能会根据案件的某个显著特征(如某个证据的显著性或某个证人的背景)迅速得出结论,忽视其他证据或因素的影响,导致决策的片面性。

正因为人类大脑处理多条信息的能力存在固有的局限,认知过载成为影响司法推理效果与效率的关键问题之一。评估大量相互交织的证据需要法官在保证决策客观性和公平性的同时,处理好多方证据的复杂关系和冲突。证据数量的繁多与证据之间的矛盾,又使得法官必须花费更多时间和精力在每一项证据的可靠性、相关性之间进行权衡。当案件事实发生变化、证词冲突或法律知识不断更新时,法官的认知负担更是加重。我国长期面临"案多人少"这一司法结构性矛盾,基层法院法官往往面临超负荷办案压力,认知资源的紧张凸显,进一步加剧了决策偏差与误判风险。在这种情况下,法官无法在一定时间内有效处理大量复杂信息,导致决策出现偏差、失误,进而影响法律的公正性与一致性。基于生成式人工智能的司法辅助系统,能够减轻法官的认知负担,有利于良好推理效果的产生,也能够促进决策效率的提升。

(二)抵消偏见驱动

偏见驱动在司法决策中的影响非常显著,尤其是当法官的主观倾向无意识地影响其对案件的判断时,偏见驱动可能产生高于认知过载的影响。[8]科学证据表明,法官与陪审员或普通民众一样,容易受到认知偏差和社会偏见的影响。前者指某种普遍的错误推理方式,后者指基于刻板印象的推理方式。[9]

法官的偏见可能会降低裁决的准确性。例如,法官在裁判时,不仅受到检察官初步要求的影响,还受到与案件无关的随机因素的影响。[10]在刑事调查场景中,无关的背景信息可能影响定罪率,因确认偏见让法官倾向于偏好有罪的调查结果。[11]类似地,对被告人的预审拘留也可能导致法官对他们作出有罪的评估。[12]在处理关于子女监护、搬迁、就业歧视的案件时,法官的决策还受到当事人性别的影响。[13]可以看出,法官在作出判决时,往往会受到过往经验、文化背景、性别认知、社会环境等因素的潜移默化的影响,而这些因素在很大程度上会扭曲他们对案件事实的理解和评估。社会偏见不仅会影响个案审理的公正,还会导致法律判决的不一致和不公平。[14]因此,结构化的决策支持系统成为弥补司法偏差、确保公平与公正的重要工具。

锚定效应是最为常见的认知偏差之一。锚定效应源于认知神经科学中的首因效应,指个体在面对决策时,过度依赖最初接触到的信息,造成这一信息在后续决策过程中对判断产生过度的影响。[15]具体来说,法官在判断案件事实时,首个证据或信息通常会对后续的证据评估产生重要影响,甚至形成固定的认知框架和路径依赖。这样的认知框架会影响法官对后续证据的解读和评估,忽视部分证据的重要性。举例来说,如果某个证据在案件初期具有显著性,那么法官可能在整个审理过程中都受到这个信息的影响,进而忽视或低估其他证据,影响判决的公正。

另一个常见的认知偏差是确认偏差,它与大脑的认知失调理论密切相关。根据这一理论,当法官在审理案件时形成了初步心证,即某个立场或观点后,大脑中的奖赏回路会驱使法官倾向于寻找与其初步心证一致的证据,从而增加支持性证据的权重。[16]上述偏向性反应会导致法官更加关注与其先入为主的观念相一致的信息,同时抑制对矛盾信息或反驳性信息的深度考量,进而使法官在评估证据时,忽视那些可能挑战其先入为主意见的重要信息。这种心理机制在司法裁决中非常普遍,尤其是在涉及复杂社会议题的案件中,确认偏差可能导致判决的严重不公。

人类的预测往往是"嘈杂的"。在输入相同信息的情况下,不同的人,甚至是同一个人在不同时间,会作出截然不同的预测。使用统计公式可以从决策中去除这种"噪音"或不相关因素。[17]因此,经过精心设计和测试的自动化系统能够有效控制或抵消偏见驱动的影响,使决策过程更加公正。[18]一方面,基于人工智能的决策支持工具旨在通过系统化、结构化的方式帮助法官克服认知偏差对司法决策的影响;另一方面,人工智能技术可以提供大量的数据分析结果,帮助法官从多个角度评估案件,识别潜在的偏

见,并根据事实和法律条文作出更加公正与一致的裁判。基于生成式人工智能的司法辅助系统,能够有效避免法官在处理证据时受到不必要的主观影响,从而减少认知偏差带来的负面效果。

## 三、信任危机:DeepSeek 类生成式人工智能嵌入司法审判的法律风险

在当今技术日益渗透司法领域的背景下,DeepSeek 类生成式人工智能的应用引发的法律风险大致可划分为“表层风险”与“深层风险”两类。前者主要源于模型在训练、调用或适配过程中的技术短板问题,属于可通过优化算法克服的阶段性问题。后者则是由于 DeepSeek 类生成式人工智能算法本身特点产生的固有弊端,是当前阶段难以通过技术改进予以克服的风险,主要包括算法歧视的继承、“黑箱”特性对决策可解释性空间的压缩和决策责任归属的模糊这三个方面。首先,DeepSeek 类生成式人工智能在司法实践中的使用会继承和放大历史数据中的歧视性偏见,导致判决加剧社会的不平等。其次,DeepSeek 类生成式人工智能的“黑箱”特性使决策过程变得不透明,严重削弱法律系统的可解释性,进而影响公众对司法公正的信任。最后,随着 DeepSeek 类生成式人工智能在司法领域的广泛应用,决策责任的归属问题亟待解决。当人工智能的预测或建议出现错误或偏见时,责任应归属于谁? 是法官、人工智能开发者还是其他相关方? 因此,在 DeepSeek 类生成式人工智能技术与司法决策结合的过程中,如何平衡技术使用与法律公正、透明以及责任归属的问题,成为当前生成式人工智能嵌入司法审判的核心议题。

### (一)继承数据歧视

近年来,可收集的数据呈指数增长。[19]然而数据收集速度的快速增加,又产生各种非标准化来源的、不完整、非结构化且混乱的数据集。尽管数据汇总形成的大数据集为分析提供了宝贵的资源,但用于决策的数据集存在收集不完整、继承偏见或缺乏背景信息等问题,容易生成误导性的结论。基于历史数据作出决策的系统,自然会继承过去的歧视。[20]人工智能系统的应用效果取决于其训练数据的质量,如果这些数据存在歧视,无论这种歧视是有意设计的还是在无意中产生的,都可能会被“固化”到人工智能的输出结果中。[21]因此,算法的“客观性”是表象,实质上可能存在对既有社会结构和历史偏见的延续与放大。当人工智能以“数据驱动”的方式参与量刑、假释、移民审查等重大决策中时,其所依赖的数据往往已经具有种族、性别、阶层等不平等结构。此时,人工智能系统作为中立司法工具的意义将被削弱。技术偏见一旦被嵌入司法流程,不仅会强化原有的社会壁垒,还可能通过自动化与规模化手段,使对弱势群体的歧视变得更加隐蔽而难以纠正。因此,在司法裁量引入人工智能技术的过程中,不仅要关注算法的性能与效率,更需要从法理、公平和人权保障的角度,反思技术治理的边界与责任。

在刑事司法系统中,Northpointe 公司开发的罪犯矫正替代性制裁分析管理系统(Correctional Offender Management Profiling for Alternative Sanctions,以下简称 COMPAS)的风险评估算法是技术应用加剧系统性歧视的典型例证。该系统声称通过分析犯罪历史、社会经济背景和社区数据预测再犯风险,却被美国非营利新闻组织“为了公民”(ProPublica)调查揭露出其存在对黑人群体的显著种族偏见:在佛罗里达州,具有相似犯罪记录的黑人被告被标记为“高风险”的比例(45%)远高于白人被告的比例(24%)。该算法的核心机制在于高度依赖种族相关的代理变量,比如邮政编码(贫困社区被默认为高犯罪风险区域)、就业状况(低收入与少数族裔群体存在统计关联)等。[22]历史上的居住隔离、就业歧视等体现结构性不公的变量可能被编码为“客观风险指标”。例如,无暴力前科的黑人被告可能因居住于犯罪率较高的贫困社区而被判定为高风险,而背景相似但来自富裕社区的白人被告却被评定为低风险。这种差异化评估直接影响量刑决策,导致前者面临更长的刑期和更少的保释机会。

当技术偏见与司法实践相结合时,还会形成自我强化的歧视性循环。比如被标记为“高风险”的被告因严厉判决丧失工作、住房等社会资源,释放后再犯风险实际升高,坐实了模型的偏见性预测,形成“预测—惩罚—强化预测”的闭环。当 COMPAS 这类人工智能系统被应用于司法领域,其依据的历史

逮捕和定罪数据本身已隐含对特定群体的系统性偏见(如少数族裔社区长期面临的过度监禁和轻罪重判)。算法将这些偏见转化为量化风险评分后,又会进一步扩大法律环境中的不平等。例如,再犯风险模型对女性和经济弱势群体存在隐性歧视,通过量刑、假释决策的自动化持续加剧资源分配失衡。技术驱动的评估体系本质上是一个"自我实现的预言",其输出的风险标签不仅反映既有偏见,更通过司法干预强化对无辩护能力群体的系统性排斥,最终导致某些群体在监禁中承担不成比例的代价。[23](P254-264)

除COMPAS外,还有多个司法与执法系统已经将人工智能应用于风险预测、资源调配等任务。例如,美国移民与海关执法局"严审"(Extreme Vetting)移民筛查算法,希望利用算法和大数据技术对签证申请人的社交媒体、数字足迹、犯罪记录和财务记录等进行全面审查。但是该类移民算法的数据训练依赖于历史数据,而这些数据本身存在固有偏见,因此某些群体在移民筛查算法中会被更高概率地标定为"高风险",[24]在移民审查中处于不利地位。英国"危害评估风险工具"(Harm Assessment Risk Tool)与COMPAS的风险评估算法相似,旨在通过分析犯罪历史、年龄、性别、邮政编码等数据,预测犯罪嫌疑人未来两年内再犯的可能性。但其数据采集来源同样包含涉及社会经济标签的数据,例如邮政区、收入、受教育程度等,可能会将贫困或边缘化群体自动打上"高风险"标签,加剧对贫困地区居民的歧视。美国"战略对象清单"(Strategic Subject List)、荷兰"系统防线提示"(System Risk Indication)也存在类似问题。由此可见,尽管人工智能以客观中立的形式嵌入司法决策当中,但决策数据本身存在的歧视性可能会通过算法"继承"下来,最终对某些群体产生结构性的不利结果。

(二)"黑箱"问题压缩解释空间

人工智能系统,尤其是机器学习算法,在其运行过程中往往存在"黑箱"特性。简言之,由于深度学习算法的层次化处理单元,算法的内部结构几乎无法被解码。人工智能系统通常以一种隐藏其决策过程的方式运作,甚至连开发者自己也难以解释其决策依据。只有输入和输出数据可被观察,数据处理过程本身在"黑暗"中进行,这被称为人工智能的"黑箱"问题。[25]

而司法实践面临的"黑箱"问题,不仅是技术可解释性的难题,更是关涉法律正义、程序透明与权利保障的挑战。在法治的语境下,可解释性不仅是裁决合理性的体现,更是监督与救济机制得以运行的前提。当司法权部分依赖于不透明、不可解释的机器输出时,可能会产生责任归属模糊、纠错机制失灵的风险。尽管人工智能系统能够根据输入的数据作出统计意义上可靠的预测和决策,但最终用户——无论是法官、律师,还是当事人,往往无法理解和解释这些决策是如何得出的,或者哪些因素在决策过程中起到了关键作用。[26]因为算法高度复杂的非线性计算缺乏透明的逻辑链条,所以用户很难解释其推理过程。法律决策可能会对个人的生活产生无法挽回的影响,直接影响到个人的权利,所以透明度的缺失在法律环境中尤为不可接受。透明度缺失会导致公众质疑法律系统的完整性,丧失对法律系统的信任,并对责任归属产生疑问。[27](P35)国产大语言模型DeepSeek的"深度思考模式"可以模拟人类多轮思考的过程,这似乎可以让用户理解DeepSeek在决策时是"怎么想"的,让其决策变得"可解释"。然而,这种"可解释性"只是复现了一定的逻辑链条,不能完全解释内部计算过程,其本质是一种"类可解释"。因此,"黑箱"带来的压缩解释空间的困境仍是无法回避的问题。此外,机器学习的不透明性还源于其他因素。例如,技术公司有动力故意隐藏源代码和相关测试数据,因为这些技术是其商业机密。又如,用户(法院或刑事诉讼中的当事人)缺乏必要的知识或技能对机器学习过程进行检查。[28]

DeepSeek类生成式人工智能算法的不透明性会对司法运行产生很大影响。在实践中,刑事司法系统在保释、判决和假释等情境中日益广泛且频繁地使用预测性算法。例如,在保释审理中,法官会借助算法结果评估个体是否有可能按时返回法庭接受审判,以及如果他们没有在审判前被拘留,是否可能再次犯罪。[29]又如,在作出判决时,法官要考虑算法结果以评估个体如果在一定期限后被释放,是否可能重新犯罪。[30]尽管基于大数据的算法结果可以帮助决策者避免依赖直觉和个人偏见进行决策,[31]从而

有效帮助法官进行预测、作出裁判,但因算法的结构、内容和测试不透明,算法辅助司法决策的工作方式仍受到强烈的抨击。[32]当人工智能决策过程不透明时,法官、辩护律师、检察官无法完全理解或预测建议是如何得出的,自然也无法对当事人进行解释和说明。[33]在这种情况下,人们很难对法律的运作保持信任。

在法律领域,算法的透明性和可解释性是确保决策者承担责任的关键工具。以法官为例,法官必须对其裁决提供合理的解释,以便上级法院审查。全球各地的法院普遍遵循这一要求,这也凸显了解释在维护法律裁决的公正性和指导未来决策方面的作用。[34]然而,若法官的裁决依赖于人工智能系统的预测结果,但该系统的决策过程不透明,法官就难以进行有效的解释,无法发现人工智能决策的错误和偏见。因此,在司法决策中引入人工智能后,确保算法决策过程透明、可解释且可以接受审查至关重要。无论作出决策的主体是人类还是人工智能系统,法律对解释的要求都是减少偏误的基本保障,也是维护司法决策过程完整性和公信力的关键。所以,在人工智能与司法应用的交汇处,“可解释性”不仅是技术问题,也是规范性难题。唯有缓解“黑箱”问题压缩司法解释空间所带来的信任危机,才能真正实现技术对法治的融入。

(三)决策责任归属不明

当 DeepSeek 类生成式人工智能系统参与保释判断、量刑建议,甚至裁决建议中时,其输出结果往往是多重决策路径交织的产物,难以用传统意义上的“单一责任人”规则进行归责。这不仅在技术上制造了模糊边界,更在制度上削弱了司法责任的清晰性。在最简单的模型中,当一个人作出决定并付诸行动时,通常认为这一主体应对自己的行为和决定承担责任。然而,责任的归属并非总是明确的。在更复杂的现实情况下,责任归属规则需要回应一系列问题:如何才能合理地界定责任?责任应该归属于谁?什么时候归属责任才是有意义的?这些问题同样适用于人工智能的领域。[35]

当法官依赖人工智能生成的风险评估结果或量刑建议时,就会引发一个问题:如果出现决策错误,究竟该由法官、人工智能开发者还是训练该算法的数据科学家承担责任?人工智能的介入会使法律裁决中的责任链复杂化,导致归责的复杂化和责任的分散化。责任的分散可能会削弱公众对司法体系的信任,因为当错误决策发生时,模糊的责任归属使确定责任方并加以纠正的过程难以实现,制约法律功能的有效发挥。

责任归属和分配的问题被称为“多手问题”,部分学者尝试通过强调多个参与者的共同责任来解决这一难题,但这一理论在司法实践中并不具备可操作性。[36]此外,为描述人工智能参与决策的责任属性,有学者提出分布式责任的概念。基于人工智能的决策或行为的效果通常是许多参与者之间无数次互动的结果,涉及设计师、开发者、用户、软件和硬件,随着分布式代理的出现,责任也被分布开来。[37]这似乎是解决多手问题的一个良好的概念性解决方案。然而,由于参与方的贡献具有不可分割性,承认责任的分布特征并不能解决如何分配责任的实际问题。同时,简单地分配多方责任可能会面临其他挑战。首先,虽然有多方参与其中,但某方可能比其他方更负责。分布式责任并不意味着责任总是应当平等分配。其次,一方或多方可能有意曲解他们的贡献,试图逃避责任。例如,在飞机或自动驾驶汽车事故中,各方可能基于不同的案情、利益作出针对性解释或辩护,这使责任归属的确定在实践中难以进行。最后,各方主体在责任链条上出现的时机亦会导致不同的责任程度,值得更细致地加权评估。[38]

## 四、重塑信任:构建“人机协同”的新型审判范式

重塑司法领域对 DeepSeek 类生成式人工智能工具的信任是司法系统适应新时代需求的内在要求。为降低该类人工智能在司法审判中的法律风险,需要构建“人机协同”的新型审判范式。

(一)明确使用边界

欧盟《人工智能法案》明确将“旨在被司法机关或代表司法机关使用的人工智能系统,以协助司法机

关研究和解释事实及法律,并将法律应用于具体事实,或在类似的替代性争议解决中使用的系统”视为“高风险”系统。与之对照的是,“旨在检测决策模式或偏离先前决策模式的系统,且不打算取代或影响之前完成的人类评估,且没有适当的人类审查的系统”则被豁免,不被视为具有高风险。该法案强调了在司法领域使用人工智能时,需要明确使用边界,确保合理的人类审查机制,避免人工智能带来的潜在法律风险。有学者认为,根据组织发展理论的原则,司法系统已经进入由人工智能重新定义问题认知、政策和制度背景,关涉身份认同的转变以及司法体系的重大变革的第三阶变革阶段。[39]例如,使用人工智能进行判决辅助决策,从根本上质疑了“应当由人类独立作出裁决”的传统观念。[40]我国最高人民法院司法改革领导小组办公室负责人曾明确表示:“在中国法院,人工智能可以辅助法官办案,但在任何情况下都不能代替法官裁判。”那么,如何明确DeepSeek类人工智能工具的使用边界,避免该类人工智能嵌入司法审判后产生的天然法律风险,成为亟待解决的问题。

根据法律支柱理论,法律体系依赖于逻辑与道德两大根基。[41]前者源于人类大脑皮层理性思维对客观规律的认知与推演能力,后者则植根于边缘系统情感体验所形成的社会共同价值准则。在司法审判领域,逻辑的具象化体现为法律规则的明确性、程序正义的严谨性与事实认定的精确性,而道德则投射为裁判者基于社会伦理对个案特殊性的衡平考量与价值判断。[42]人工智能作为逻辑运算的极致产物,通过算法模型对海量司法数据进行结构化处理,能够实现法律条文与案件事实的精准匹配、诉讼流程的自动化推进以及裁判结果的概率化预测。这种技术特性决定了人工智能在司法场景中具有高效处理标准化、程式化案件的先天优势。法律领域的特点与技术领域有很大不同,因为法律不仅基于逻辑,还与情感紧密相连,如正义感、生活中的舒适感、商业连续性、家庭团结和亲情等。在民事案件中,法官通常是被动的,但他们必须采取更具说服力的方式来调解当事人纠纷,这一过程建立在互助合作的基础上。法官能够以富有同理心和关怀的方式处理情况,这对于法庭的公平至关重要,同时也是实现和解并确保有效解决问题的关键。[43]一般而言,人工智能常用于简化那些耗时多且风险较低、不太可能显著改变结果的日常任务。但随着人工智能越来越多地被用于解决复杂问题,简化引发的歧视风险也随之增加。在处理多维问题时应用标准化解决方案,往往会简化这些问题的不确定性。人类在识别自动化无法捕捉的细微差别方面表现更加出色,这凸显出人工智能在处理依赖于特定语境的复杂决策方面与人类存在显著差距。[44]

因此,将人工智能系统深度融入现代司法体系的首要路径,在于构建科学合理的案件分类机制。根据上述的法律支柱理论,刑事案件和民事案件之间有着显著的区别。在刑事案件中,逻辑因素比道德因素更重要,因为案件审理的重点往往是基于标准确定事实是否符合特定法定事实要件。比如,上海刑事案件智能辅助办案系统(206系统)便聚焦于刑事审判中程序性与逻辑性最强的环节,如刑事案件中的证据标准指引、单一证据校验、逮捕条件审查、社会危险性评估、证据链和全案证据审查判断等,体现出将高规范度任务与人工智能相结合的应用思路。[45]而在涉及婚姻、继承、家庭和经济事务的民事案件中,更注重道德因素的考量,因为其目标是通过法庭中的谈判和调解解决问题。进一步而言,还可将民事案件审理工作精准划分为“行政性事务”与“裁量性事务”两大类型。前者涵盖证据形式审查、诉讼时效计算、格式合同条款效力判定、程序性文书生成等具有高度重复性与确定性的司法辅助工作。此类事务本质上属于法律逻辑的机械性适用,可通过自然语言处理技术与知识图谱系统实现全流程自动化。后者则涉及过错责任的比例划分、公序良俗的具体适用、利益平衡的准确把握等需要价值判断的领域,必须由法官通过庭审捕捉当事人真实诉求,运用法律解释学方法在法律规制空白处进行创造性续造,以共情能力化解对抗情绪、以调解艺术修复社会关系。[43]

(二)发展可解释人工智能

面对许多人工智能系统存在的“黑箱”问题,以及这一问题在司法领域衍生出的公正性、可审查性、可解释性需要,目前有“外生方法”与“分解方法”可以帮助用户理解机器学习模型如何作出预测。“外生方法”不直接解释机器学习算法的内部运作,而是通过提供一些额外的、与算法运作无关的外部信息,帮

助用户理解模型的表现。“分解方法”则尝试从根本上解释机器学习模型如何作出决策:它不是仅提供外部信息,而是直接解释模型的推理过程,甚至尽力复制模型内部的推理步骤。① 通过这种方式,用户可以理解模型如何在不同的因素之间作出选择。[46]

可解释的人工智能主要依赖于“分解方法”,普遍使用使人工智能决策过程透明和易于理解的机制。[47]“可解释的人工智能”的定义是:在特定受众面前,可解释的人工智能应当提供详细信息或理由,使其运行方式清晰易懂。[48]在司法系统中,这一点尤为重要,因为它允许法官、律师和其他相关方理解人工智能如何得出结论。可解释的人工智能对人工智能决策过程的揭示,使得可能被忽视的偏见或错误更容易被发现并纠正。[49]

同时,可解释的人工智能可以增强人类与机器之间的协作决策过程。在司法审判过程中,人工智能系统可以通过提供基于数据的洞察和初步分析协助法官,而最终的决策权仍然属于法官。这样的合作确保人工智能支持司法职能,而不取代法官所具备的细致判断和伦理考量。通过过程透明,可解释的人工智能使法官能够理解并信任人工智能提供的洞察,从而作出更为知情的决策。可解释的人工智能在建立公众对司法系统的信任方面也发挥着关键作用。在我国当前法治环境中,公众对人工智能参与司法裁判的信任尚未完全建立。尤其在涉及人身安全、财产安全的案件中,人民群众普遍要求“看得见的正义”,即不仅结果公正,更要求过程透明、理由可理解,可解释的人工智能将有助于司法机关回应公众对人工智能辅助决策过程透明的现实关切,让公众对人工智能驱动的决策的公正性和透明性产生信心。人工智能系统对其推理过程的解释有助于向公众揭示技术的本质,使人们更容易接受并信任人工智能在司法过程中的使用,维持法律系统在公众眼中的完整性和可信度。

尽管可解释的人工智能具有诸多优势,但其实施也面临挑战。既准确又可解释的人工智能系统开发在技术上非常复杂,且具有资源密集特性。通常,人工智能模型的准确性与可解释性之间存在权衡关系。更复杂的模型,如深度学习算法,通常透明性较差但准确性较高;而较简单的模型虽然更易解释,但可能牺牲准确性。因此,平衡这些因素对于可解释的人工智能在司法决策中的成功应用至关重要。

(三)确立法官主体的决策责任

责任的归属依赖于责任产生的条件。“控制”条件与“认知”条件被认为是评定个体是否应该为某个行为负责的标准。[50]“控制”条件意味着责任归属必须考虑到个体是否对自己的行为有足够的控制力。[51](P98)“认知”条件要求主体在行动时知道并意识到自己正在做什么。正如亚里士多德所指出的,任何行为的产生必须来自于代理人,且代理人必须知道自己在做什么,否则不能视其为有责任的人。[52](P23)反之,若主体在行动时对后果一无所知,或者根本不知道自己在做什么,那么不应为该行为承担责任。人工智能无法真正“自由地行动”,其既不具有自由意志,也并不知道自己在做什么。因此,唯一可行的选择是由人类为有人工智能技术参与的行为承担责任。[53]即使某些人工智能具有行动或决策的能力,但由于道德代理能力的缺失,其行为或决策也应由开发和使用这些技术的人类负责。当前法律体系在处理涉及非人类行为时,通常采取的做法就是如此。

明确承担责任的主体为人类后,在司法审判中如何确定责任归属值得进一步探讨。责任归属问题和责任分配问题具有时间维度:谁在什么时间(和什么地点)做了什么?在技术的使用和开发过程中,通常会涉及长时间跨度的人类代理链。就人工智能而言,情况尤为如此,因为复杂的软件通常有着长时间的开发历程,涉及许多开发者在不同阶段为不同部分的软件作出的贡献,软件开发任务可能在同一组织内部或不同组织间流转。对于机器学习人工智能,还涉及生产、选择、处理数据和数据集的过程。人工智能软件也可能最初为某个应用场景开发,但后来被用在完全不同的应用场景中。此外,在技术的使用和开发流程之外,还涉及防止人工智能系统发生故障的维护程序。因此,必须检查数据处理与人工智能

①这里的“解释”概念指的是提供对算法内部状态的洞察,或展示人类可以理解的算法近似值。

开发、应用、维护等各个环节,才能明确责任归属。但是在人工智能迅速迭代的情况下,进行这样的调查变得非常困难,将司法实务中的担责主体界定为人工智能开发者或数据训练者将付出极大的社会成本。

负责任不仅是指对自己的行为负责,还涉及对"受影响的对象"负责,即作为使用者的用户对受影响者具有回应义务。"受影响的对象"可以是被代理人行为直接影响的人,也可以是其他间接受到影响的人。[54]无论是谁,只要受到影响,就有权要求代理人解释自己的行为。比如,公众有权要求法官解释其判决。欧盟《一般数据保护条例》规定了"解释权",意味着个人有权知道自动化系统如何作出影响他们的决策。[55]2014 年,党的十八届四中全会提出"谁办案谁负责,谁签字谁负责"的审判责任理念,确认实行办案质量终身负责制和错案责任倒查问责制。法官若因故意或重大过失导致裁判结果错误并造成严重后果,可能会面临停职、延期晋升、调离审判执行岗位、退出员额等惩戒,或依据《中华人民共和国公务员法》《中华人民共和国法官法》《中华人民共和国公职人员政务处分法》《人民法院工作人员处分条例》等规定受到处分。人工智能所生成的输出或分析结果,极大地依赖于输入数据的数量和质量。人工智能的缺陷,如生成不准确或幻觉性的输出,必须通过将结果与准确、可靠、事实性的数据进行比对来预见和解决。在此语境下,法官必须承担仔细验证和交叉检查结果细节的责任。若法官未加核实就直接使用人工智能输出结果中存在误导信息或错误信息,便可能违反审判职责。因此,无论人工智能工具是否具有技术自主性,也无论人工智能在设计与使用过程中涉及的主体有多少,其在司法活动中的应用都是在法官理性支配下进行的,因此将法官确立为使用 DeepSeek 类生成式人工智能的责任主体是应有之义。同时,鉴于人工智能的"黑箱"和不可追溯性两大特征,也难以将司法过程中使用 DeepSeek 类生成式人工智能的责任主体界定为除法官外的其他主体(如人工智能开发者等)。

## 五、结语

近年来,我国法院系统一直以积极主动的姿态拥抱数字化变革,通过一系列举措积极促进司法与科技深度融合发展。DeepSeek 类生成式人工智能具有提升司法效率和保障司法公正的潜力,但使用这一类大语言模型时会放大其本身固有的数据歧视与"黑箱"问题,加之人工智能介入司法审判带来的责任归属问题,导致信任危机的出现。为重塑信任,需要构建"人机协同"的新型审判范式。对 DeepSeek 类生成式人工智能参与司法过程的边界予以明确,将其使用限制于逻辑判断,而非价值判断;在发展可解释人工智能的同时,确定法官为责任归属主体。DeepSeek 类生成式人工智能应作为工具辅佐司法工作,而非取代人类的裁决权与责任。通过明确边界、提升可解释性、加强责任归属,可以在充分利用人工智能的高效协同决策的同时,确保法律公正与透明,进而提升公众对司法系统的信任度。虽然法官的角色无法被人工智能取代,但是人工智能的使用不可避免地会对法官产生负面影响,导致他们产生依赖,减少寻求原始文献的动力,甚至没有动力作出符合道德或伦理标准的决策。对于该负面影响的规制,未来仍需要进一步展开探究。

(本研究为山东大学法学院研究生科研创新项目"新质生产力下知识产权保护我国数字产业安全的体系构建"资助项目。)

**参考文献:**

[1] Germain T. Judges given the ok to use ChatGPT in legal rulings[EB]. https://gizmodo.com/uk-judges-now-permitted-use-chatgpt-in-legal-rulings-1851093046,2023-12-12/2025-05-10.

[2] 樊传明. 被敲视的法官:数字司法对审判权运行的影响[J]. 法制与社会发展,2024,30(3):137-153.

[3] 刘叶萍,袁小群. 认知神经科学视角下的数字阅读认知机制研究进展[J]. 图书情报知识,2023,40(6):129-139.

[4] Pessoa L. Understanding brain networks and brain organization[J]. Physics of Life Reviews,2014,11(3):400-435.

[5] Miller G A. The magical number seven,plus or minus two:some limits on our capacity for processing information[J]. Psychological

Review,1956,63(2):81-97.
[6] Kannengiesser U,Gero J S. Design thinking,fast and slow:a framework for Kahneman's dual-system theory in design[J]. Design Science,2019,5:1-21.
[7] Guthrie C,Rachlinski J J,Wistrich A J. Blinking on the bench:how judges decide cases[J]. Cornell Law Review,2007,93:1-43.
[8] 李安. 司法过程的直觉及其偏差控制[J]. 中国社会科学,2013,(5):142-161,207-208.
[9] Zenker F. De-biasing legal factfinders[A]. Stein A,Tuzet G. Philosophical foundations of evidence law[M]. Oxford:Oxford University Press,2021:395-410.
[10] Bystranowski P,Janik B,Próchnicki M. Anchoring effect in legal decision-making:a meta-analysis[J]. Law and Human Behavior,2021,45(1):1-23.
[11] Rassin E. Context effect and confirmation bias in criminal fact finding[J]. Legal and Criminological Psychology,2020,25(2):80-89.
[12] Lidén M,Gräns M,Juslin P. "Guilty,no doubt":detention provoking confirmation bias in judges' guilt assessments and debiasing techniques[J]. Psychology,Crime & Law,2019,25(3):219-247.
[13] Miller A. Expertise fails to attenuate gendered biases in judicial decision making[J]. Social Psychological and Personality Science,2019,10(2):227-234.
[14] 蔡艺生. 论经验隶属性和法治化程度对司法裁判的影响——基于认知科学角度的实证分析[J]. 社会科学研究,2017,(3):55-61.
[15] 杨彪. 司法认知偏差与量化裁判中的锚定效应[J]. 中国法学,2017,(6):240-261.
[16] Peters U. What is the function of confirmation bias? [J]. Erkenntnis,2022,87(3):1351-1376.
[17] Berkooz G. Reducing noise in decision making:interaction[J]. Harvard Business Review,2016,94(12):18-19.
[18] Jung J. Simple rules for complex decisions[EB]. https://arxiv.org/abs/1702.04690,2017-04-02/2025-05-10.
[19] Gioia G. Artificial intelligence (AI) and judicial independence:balancing transparency and control[J]. Trauma and Memory,2025,12(3):100-114.
[20] Zemel R,Wu Y,Swersky K. Learning fair representations[A]. International conforence on machine learning[C]. Atlanta:PMLR Workshop and Conference Proceedings,2013,28(3):325-333.
[21] Verma S. Weapons of math destruction:how big data increases inequality and threatens democracy[J]. Vikalpa,2019,44(2):97-98.
[22] Mateen H. Weapons of math destruction:how big data increases inequality and threatens democracy[J]. Berkeley Journal of Employment and Labor Law,2018,39(1):285-292.
[23] Angwin J. Ethics of data and analytics[M]. New York:Auerbach Publications,2022.
[24] 王中原. 算法规制型国家:国家监管的智能转型及其政治影响[J]. 公共治理研究,2025,37(3):19-38.
[25] 郭全中,李黎. 生成式人工智能将通向隐秘的社会? ——一个叠合黑箱的逻辑与实践[J]. 暨南学报(哲学社会科学版),2024,46(12):81-96.
[26] 杨永兴. Deepseek 等开源模型法律风险治理研究[J]. 四川轻化工大学学报(社会科学版),2025,41(4):27-38.
[27] Pasquale F. The black box society:the secret algorithms that control money and information[M]. Cambridge:Harvard University Press,2015.
[28] Hildebrandt M. Algorithmic regulation and the rule of law[J]. Philosophical Transactions of the Royal Society A:Mathematical,Physical and Engineering Sciences,2018,376(2128):20170355.
[29] 刘宇琪,秦宗文. 刑事证明中的预测性算法证据研究[J]. 中国人民公安大学学报(社会科学版),2024,40(1):75-88.
[30] Wiseman S R. Fixing bail[J]. The George Washington Law Review,2016,84:417-426.
[31] Deeks A. The judicial demand for explainable artificial intelligence[J]. Columbia Law Review,2019,119(7):1829-1850.
[32] Roth A. Trial by machine[J]. The Georgetown Law Journal,2015,104(3):1245-1303.
[33] Liu H W,Lin C F,Chen Y J. Beyond State v Loomis:artificial intelligence,government algorithmization and accountability[J]. International Journal of Law and Information Technology,2019,27(2):122-141.
[34] Eidelson B. Reasoned explanation and political accountability in the roberts court[J]. The Yale Law Journal,2021,130(7):1748-1826.
[35] 林洹民. 论人工智能致损的特殊侵权责任规则[J]. 中外法学,2025,37(2):344-362.
[36] Van de Poel I. The problem of many hands:climate change as an example[J]. Science and Engineering Ethics,2012,18(1):49-67.
[37] Taddeo M,Floridi L. How AI can be a force for good[J]. Science,2018,361(6404):751-752.
[38] Coeckelbergh M. Artificial intelligence,responsibility attribution,and a relational justification of explainability[J]. Science and Engineering Ethics,2020,26(4):2051-2068.
[39] Castelnovo W,Sorrentino M. The nodality disconnect of data-driven government[J]. Admin. & Society,2021,53(9):1418-1442.

[40] Dhungel A K. "This verdict was created with the help of generative AI…?" on the use of large language models by judges[J]. Digital Government:Research and Practice,2025,6(1):1-8.
[41] Habermas J. Law and morality[A]. Hill T E, Scalia A, Said E W,et al. The tanner lectures on human values[C]. Salt Lake City:University of Utah Press,1988:217-279.
[42] 崔立红,李昶彬. 总体国家安全观视阈下专利安全理念的实现路径研究[J]. 科技进步与对策,2024,41(23):19-28.
[43] Tampubolon Y S H,Murwadji T. The application of artificial intelligence in civil trials:mechanism vs. humanism[J]. Journal of Ecohumanism,2025,4(1):1339-1352.
[44] Akash K. Improving human-machine collaboration through transparency-based feedback-part I:human trust and workload model [J]. IFAC-PapersOnLine,2019,51(34):315-321.
[45] 詹建红,邱宇欣. 人工智能嵌入侦查讯问的应用风险及其制度应对[J]. 浙江大学学报(人文社会科学版),2025,55(5):112-127.
[46] Edwards L,Veale M. Slave to the algorithm? why a "right to an explanation" is probably not the remedy you are looking for[J]. Duke Law & Technology Review,2017,16:18-84.
[47] 杨志航. 算法透明实现的另一种可能:可解释人工智能[J]. 行政法学研究,2024,(3):154-163.
[48] Dwivedi R. Explainable AI (XAI):core ideas,techniques,and solutions[J]. ACM Computing Surveys,2023,55(9):1-33.
[49] Górski Ł,Ramakrishna S. Explainable artificial intelligence,lawyer's perspective[A]. Proceedings of the eighteenth international conference on artificial intelligence and law[C]. Burlington:MorganKaufmann Publishers,2021:60-68.
[50] Rudy-Hiller F. The epistemic condition for moral responsibility[EB]. https://plato. stanford. edu/entries/moral-responsibility-epistemic/? fbclid=IwAR0N1LukDRwztd9PYvtm8jlqhS8EStGCfEnKne1JOPstML5iEZN8arsa8Sc,2018-09-12/2025-05-10.
[51] Fischer J M,Ravizza M. Responsibility and control:a theory of moral responsibility[M]. Cambridge:Cambridge University Press,1998.
[52] Aristotle. Nicomachean ethics[M]. London:ReadHowYouWant,2006.
[53] 郑海蓉. 人工智能致损的责任主体认定与责任分担[J]. 青岛科技大学学报(社会科学版),2025,41(1):70-78.
[54] Duff R A. Who is responsible,for what,to whom[J]. Ohio State Journal of Criminal Law,2004,2:441-461.
[55] Metikoš L,Ausloos J. The right to an explanation in practice:insights from case law for the GDPR and the AI act[J]. Law,Innovation and Technology,2025,17(1):205-240.

# Reshaping Trust: The Underlying Logic and Optimization Path of Embedding DeepSeek-like Generative AI in Judicial Adjudication

Cui Lihong Li Changchen
(School of Law, Shandong University, Qingdao 266237, China)

**Abstract**: The application of generative artificial intelligence (AI) like DeepSeek in judicial proceedings can alleviate judges' cognitive burden, prevent cognitive overload, and enhance judicial efficiency. Meanwhile, it can counteract judicial cognitive biases, avoid bias-driven outcomes, and ensure judicial fairness. Nevertheless, trust remains a primary barrier to embedding DeepSeek-like generative AI in judicial trials, manifesting through issues such as inherited data discrimination, the "black box" problem constraining interpretability, and ambiguous liability for AI-assisted decision-making. To rebuild trust in such technology within judicial practice, a paradigm shift toward "human-AI collaboration" must be achieved through three paths: (1) clarifying the boundaries for deploying generative AI like DeepSeek; (2) advancing explainable AI (XAI); and (3) affirming judges' ultimate decision-making responsibility.

**Key words**: DeepSeek; explainable artificial intelligence (XAI); cognitive psychology; decision-making liability; human-AI collaboration

责任编辑:王明舜