

“能—智分合”：司法 AI 的分阶段发展模式

孙笑侠 魏义铭*

摘要 在司法人工智能明确定位于“辅助”后,未来中国司法人工智能的发展方向仍应是提升辅助型司法 AI 之“智”。借鉴“德雷福斯模式”,基于 AI“能”与“智”二维可分离关系,我国可以建构辅助型司法 AI“能—智分合”模式。当前提升司法 AI“智”存在因果判断和价值判断的技术瓶颈,但并非没有技术解决方案。我国辅助型司法 AI“智”的提升可通过三种技术方案分阶段实现:一是“与法官一起思考”,以可解释性为抓手,在形式层面实现可用性的突破;二是“与法官的思考对齐”,在法律专业技能与职业伦理的约束下推进价值对齐;三是“像法官一样思考”,攀登因果判断和价值判断的阶梯,在“辅助”定位下增强对复杂裁量问题的推理支撑。

关键词 司法人工智能 辅助型司法 AI 德雷福斯模式 可解释性 价值对齐

一、引言

2022 年最高人民法院在《关于规范和加强人工智能司法应用的意见》(以下简称《意见》)中明确规定了人工智能的“辅助审判原则”——“坚持对审判工作的辅助性定位和用户自主决策权,无论技术发展到何种水平,人工智能都不得代替法官裁判。”^{〔1〕}该《意见》明确叫停了“决策型司法 AI”,还规定了五类应用范围(第 8—12 条)。这意味着,我国已经明确宣布司法人工智能是有“止境”^{〔2〕}的,并将其定位于对司法的辅助性功能。这在技术认知与司法认知两方面都是相当明智的,也是进步的。为了行文简便,本文把这种具有辅助性的司法人工智能简称为“辅助型司法 AI”。

* 孙笑侠,浙江大学光华法学院教授、法学博士;魏义铭,复旦大学法学院博士研究生。本文系教育部人文社会科学重点研究基地重大项目(2020—2024)“科技与人权的关系研究”(项目编号:20JD820004)的阶段性成果。

〔1〕 参见《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。

〔2〕 参见孙笑侠:《论司法信息化的人文“止境”》,载《法学评论》2021 年第 1 期。

从前些年智慧法院建设的全面铺开,到2022年骤然宣告AI仅限辅助性功能,这个急转可能带来一些疑惑:既然司法AI被界定为“辅助型”,是否就意味着对其智能水平的期待可以相应降低?这会不会带来预期的下调:司法机关将不再强求司法AI具备“像人那样思考”的完整心智能力?我们的基本预设是:即使司法AI被定位于辅助型,其仍然需要在具有较强“能力”的同时,持续提升“智力”水平。但是,司法AI的“智”不可能在短期内提升,因而需要分阶段推进。这也促使我们去思考一个问题:未来五至十年中国司法AI的发展方向、路径和阶段性任务应该是怎样的?

本文借鉴“德雷福斯模式”(Dreyfus model)^{〔3〕},提出司法AI“能—智分合”技术演进模式。虽然司法AI定位于辅助型,但它仍在“智”维的提升上有空间、有必要和有可能,且应当成为未来五至十年的方向和任务。司法AI的研发路径应当遵循“能”维与“智”维之间的分离—耦合规律:在确保工具性功能稳步发展的前提下,逐步推动“智”维的增长与场景适配。诚然,司法AI在“智”维的提升方面确实面临若干瓶颈与制约,但是,辅助型AI“智”维的提升仍然是有可行路径的。

二、司法AI“能—智分合”模式建构及其应用

无论是人类的智能还是人工智能,进行“能”与“智”的二维区分都具有重要的意义。区分“能”(行为)与“智”(思维),一直是我们认识AI发展的两个抓手。“能”与“智”的可分与可合关系,也是我们理解以“能”测“智”、以“能”代“智”的认识前提。图灵测试就是以“能”测“智”,典型地将人工智能的目标设定在“像人类一样行为”的路径上。^{〔4〕}“能”与“智”的分离在我国哲学界鲜有相关的探讨,直到2024年才出现以人工智能的“智”与“能”为题的专文。该文认为人工智能之“智”的分析仅仅是一个先验性的理论分析,而它能否拥有其“能”却是一个经验性的和实践操作的问题。因此,未来人工智能的研究既要考虑理论上的先验可能性,同时也要考虑实践上的经验可行性。^{〔5〕}

就司法领域而言,司法活动中有大量事务性工作需要AI辅助。而这些司法AI能否“胜任”工作的关键,并不取决于其能力的强弱,而取决于AI之“能”、AI之“智”与场景任务(事)之间能否形成稳定的适配关系。可以说司法AI风险的主要来源就在于这三方面的不匹配。相应地,避免风险的要义也在于把握三者之间的发展规律。因此,本文借鉴了“德雷福斯模式”中关于“能—智”的关系,构建司法AI“能—智分合”发展模式。

德雷福斯(Hubert Dreyfus)早在1972年发表的*What Computers Can't Do*(1992年再版)中,

〔3〕 由美国德雷福斯兄弟基于现象学与人工智能的关系提出的技能获得模型给予了我们有益的启示。该模型最初由休伯特·德雷福斯(Hubert Dreyfus)和斯图尔特·德雷福斯(Stuart Dreyfus)在20世纪80年代提出,并在*Mind over Machine*(1986)中加以系统化,强调人的技能习得过程并非简单的规则积累,而是伴随经验、直觉与情境理解的不断深化。See Hubert L. Dreyfus & Stuart E. Dreyfus, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*, Basil Blackwell, 1986.

〔4〕 See Alan M. Turing, *Computing Machinery and Intelligence*, 59 *Mind* 433 (1950). 图灵在其中提出,为避免“机器能否思考”这一哲学难题,他将该问题替换为一个可操作性更强的“模仿游戏”——如果一个机器能够在文字交流中让人类评估者无法区分它与真人,那么它就被认为通过了测试。

〔5〕 参见江海全:《“智”的可能性与“能”的极限性:论人工智能的“智”与“能”》,载《四川师范大学学报(社会科学版)》2024年第6期,第63页。

通过生物学假设、心理学假设、认识论假设和本体论假设，系统批评了当时盛行的符号主义人工智能。德雷福斯认为，人类智能并非单靠形式化的符号处理就能还原，因为理解依赖于身体经验、情境背景和非形式化的直觉判断。即便计算机能在规则封闭的领域内表现出色，这种“能”也无法实现真正的人类“智”。〔6〕在他的分析框架里，较低层次的刺激—反应式操作或简单的规则执行可以被程序模拟，而更高层次的感知、思维和情境理解则无法完全程序化，因此“能”与“智”必须区分对待。严格来说，人类的智与能也是可以分开的。如果按照“从新手到专家”(Novice to Expert)的“德雷福斯模式”，可以驾驶员为例将从“能”到“智”分阶段讨论：刚学车的人可能机械地按照教练教的步骤完成操作(会打方向盘、会换挡)，这是具备了“像司机一样行为”的能力，但还缺乏独立判断与预判的智慧。随着经验的积累，驾驶者的“能”和“智”逐渐结合。驾驶者不仅能操作车辆，还能提前预判路况、根据天气或交通流量做出最佳路线选择。再后面的阶段是高“智”的状态，熟练的司机可以在复杂环境中迅速评估多种方案，并基于全局考量做出最优决策，这相当于人工智能从“行为模仿”走向“认知自主”。

基于 AI 的能与智的分离与融合关系，参考“德雷福斯模式”，本文把 AI 的演进步骤表述为“能—智”二维分合演进模式。以这个模式来提出关于 AI 演进的步骤，人工智能的发展可以从两个维度加以刻画：其一是执行能力(capacity, 简称“能”)，即系统在外部的任务表现上的功能性与效率；其二是认知智慧(intelligence, 简称“智”)，即系统在内部实现中体现出的理解、推理与判断能力。二者之间既分又合，在相互独立和分离的状态下促进各自提升增强，又在适当时机下可耦合并轨，在技术演化中趋向融合，加速人工智能之“智”的提升。人工智能技术的演化可划分为四个典型阶段：1. 阶段 0(初始弱能弱智)；2. 阶段 1(能优先)；3. 阶段 2(智补强)；4. 阶段 3(能智融合)。以二维坐标系形式呈现(见图 1)，横轴表示“能”的水平，纵轴表示“智”的水平。AI 研发实践中将能与智分离，缘于技术探索可行性，由此我们可以预言：这二者之间的张力应该成为行业垂直大模型研发的一个核心焦点。

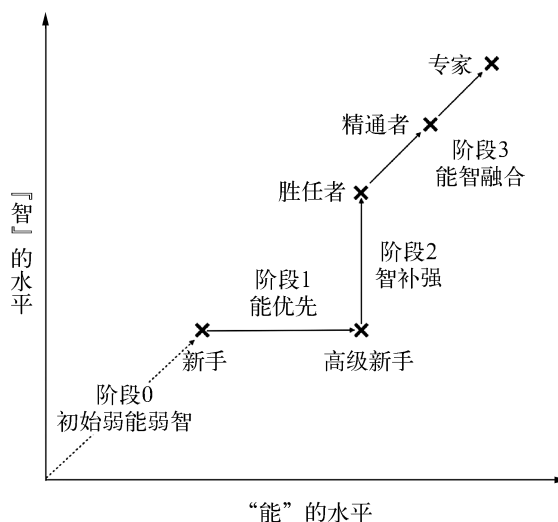


图 1 “能—智二维分合”技术演进模式

在“能—智分合”技术演进模式下，人工智能的“能”与“智”可按轻重缓急分阶段演进。落实到司法场景领域中，这一模式可转换为“辅助型司法 AI 能—智二维分合”演进模式，以下简称司法 AI“能—智分合”模式。该模式可为司法 AI 提供一套渐进、动态的成长框架，避免“盲目高能低智”或“过度依赖黑箱”。同时，该模式也可契合中国司法 AI 的发展实际，如辅助性定位、渐进升级、风险管控等。此框架可用于明确我国司法 AI 研发与应用阶段的定位、研发步骤和提升空间。

（一）阶段之定位

根据我国司法 AI 研发与应用水平，它在“能—智分合”技术演进模式中的定位，可作以下

〔6〕 See Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason*, MIT Press, 1992, p.159 - 206.

描述:

阶段0(司法人工智能的探索期),该阶段司法人工智能的特征是:无论任务的执行能力还是认知水平都处于较低的水平。典型如我国20世纪80年代开始研发的基于符号逻辑的专家系统,其任务完成率有限且难以形成真正的情境理解。^{〔7〕}又如,我国在2014—2017年期间的诸多司法应用,主要表现为基于规则引擎的单一功能系统,用于为法官提供法律条文查询、法规解释等支持功能。大部分的运用实际上并未达到开发者或使用者优先所期待的功效。在一些基层法院中,不少应用系统的使用频率并没有显著增加,某些应用系统的使用频率甚至“不升反降”。^{〔8〕}

阶段1(能优先)可以理解为司法人工智能在完成前期探索之后,依托人工智能技术的快速进步而进入的能力跃升期。与阶段0相比,这一阶段的进展,首先得益于司法信息的数字化与统一化,为系统提供了更为标准化和丰富的数据基础;其次,深度学习方法的应用,使司法人工智能在特定任务域内的表现显著提升。尤其是自2022年起大语言模型的加入进一步强化了司法人工智能的基础能力,使其在语言处理、模式识别和任务生成等方面展现出更强的潜力。这种能力的增强使得司法人工智能在部分场景中不仅能够实现稳定的输出,甚至能在一些简单的问题上超越普通法官的效率,从而体现出明显的“能”增长。然而这种提升主要体现在操作层面,而非整体认知水平的进步。司法人工智能在“智”的维度上仍显薄弱,具体表现在缺乏稳健的因果表征、可验证的解释链、自我反思与价值权衡机制。若以“德雷福斯模式”来比照,这一阶段更接近于“按情境熟练执行”的水平:模型已经能够稳定复用既有策略,但其理解能力仍主要依赖模式匹配和启发式归纳。结合我国的司法实践,自2017年以来大量应用大致处于这一阶段:法律检索在召回率与相关性方面有了显著提升,文书生成在模板化和要素回填上趋于成熟,案件执行辅助与证据分析环节通过自动化缩短了时延并降低了差错率,同时在司法管理和环节也实现了供给扩容。这些成果虽然为司法工作带来效率上的突破,但其本质仍体现为“能”的不断扩张,而非“智”的增强。

阶段2(智补强)可理解为在维持现有司法AI高“能”的前提下,有计划地扩展其“智”。这一阶段的核心任务在于突破既有“能”的局限,不再仅仅依靠效率和规模化部署的延展,而是在“能”已趋稳定的基础上,引入更复杂的推理框架与知识建模机制,使系统逐步具备情境感知、因果推理、自我解释与价值判断等更深层的能力。需要注意的是,“智”的提升并不意味着短期内必然带来“能”的同步增长。相反,因果链条的嵌入、价值权衡的引入和可验证解释机制的建立往往会提升效率成本,^{〔9〕}但这种成本付出恰恰是司法AI获得可信性与正当性的必要条件。换言之,“智”并非“能”的线性延伸,而是对其的约束与补强。若以德雷福斯的“技能习得模型”为参照,阶段2是从“高级新手”向“胜任者”乃至“精通者”过渡的阶段。这一阶段的特征不再是依赖既有规则的重

〔7〕事实上,自20世纪50年代开始,有关“法律专家系统”的研究可谓此起彼伏,梅尔在《法律世界的自动化理论》一文中提出法律信息修正理论,此后,美国于1977年开始研发在税法领域得到应用的法律专家系统(TaxMan)。伴随着法律专家系统运动在20世纪80、90年代在全世界逐渐流行,我国于20世纪80年代也开始了法律专家系统的研制。这一研制的思路最早可以上溯至龚祥瑞和李克强在1983年发表的《法律工作的计算机化》。此后,在我国出现了一大批有关电脑量刑的研究成果,但似乎都未取得明显的效果。

〔8〕参见叶燕杰:《智慧法院建设中的实践难题与破解路径——基于B市智慧司法实践的考察》,载《山东大学学报(哲学社会科学版)》2022年第3期。

〔9〕此类成本的提升可类比为:DeepSeek、ChatGPT等大模型的推理模型相较于传统模型在输出时将耗费更长的时间成本,但却可以带来更为准确的输出。

复与复用,而是对具体情境的理解和对因果关系的把握。^[10] 现有司法 AI 之“能”发展迅猛,但已处于瓶颈期。而“智补强”是我国司法 AI 在 2025 年至 2035 年两个五年规划的任务,也应当是我国司法 AI 未来十年的目标。

阶段 3(能智融合)则代表司法 AI 发展的理想状态,即执行能力与认知智慧的深度结合。在这一阶段,系统不仅能高效地完成复杂任务,还能够展现出与人类相近的理解力与判断力:司法 AI 不仅可以理解案件事实,适应新的情境,还可以在规范冲突与价值权衡中生成可解释、可验证的理由。这种“能”与“智”的统一意味着司法 AI 不再只是工具,而是复杂司法语境中真正的认知伙伴。在这一意义上,阶段 3 对应着德雷福斯模式中的“专家”乃至“实践智慧”层次:系统的运行已不再依赖于规则的逐条调用,而是能够在具体语境中进行整体把握与创造性回应。从“能”与“智”的关系来看,阶段 3 意味着二者之间张力得以化解。过去以“能”为主导的效率扩张和以“智”为导向的补强,在这一阶段逐渐汇合,形成兼具操作可行性与认知可信性的整体形态。可以说,阶段 3 是司法 AI 向“强人工智能”方向靠近的阶段愿景,其核心不在于完全复制人类的意识或主观体验,而在于在司法制度运行中实现一种高度整合的“能—智”功能体系,使技术既能回应现实案件的复杂性,又能承载法律正当性的内在要求。

(二) “智补强”之必要

司法 AI“能—智分合”框架对司法 AI 研究的理论意义在于:其一,提供了人工智能能力演化的双维度分析框架,有助于在司法 AI 研究和实践中区分“外在功能性”与“内在认知性”;其二,可为技术路线规划提供理论参照,区分司法 AI 不同发展阶段的研究重点(如在阶段 1 后应聚焦于“智”的补强,在阶段 2 应探索多模态推理与价值融合机制);其三,具有伦理与政策分析价值,高“能”低“智”的系统可能在缺乏充分理解的情况下造成不可预见的风险。因此提升“智”维度是司法 AI 技术治理的关键任务。由此可见,这个理论框架不仅能够解释现有司法 AI 的技术分布状态,还能预测其未来演化趋势,为司法 AI 研究、发展与治理提供具有指导性的理论支点。

近年来,中国司法体系在人工智能的推动下,逐步形成了一套司法 AI 辅助型系统,包括:“司法流程辅助类”“类案推送与量刑辅助类”“诉讼行为监管与司法决策评估类”“司法公开与公共服务类”“司法大数据分析 with 政策建议类”等。整体而言,越是任务清晰、规则明确的环节,越易实现自动化,体现出司法之“能”的可行性;反之,越是涉及规范解释、事实认定与价值判断的环节,越难以形式化建模,这也是司法之“智”的瓶颈所在。在司法辅助型 AI 中,分离“能”和“智”的应用价值在于:第一,明确技术边界。“能”维可量化的是数据处理、检索、归档、类案推送等可编程任务;“智”维涉及理解、推理与价值判断,尤其是在法律规范和个案情境适配上。区分二者有助于避免将价值判断完全交由机器,防止越权或误用。第二,优化资源配置。通过对“能”与“智”的二维分合评估,可在系统设计中将高频、可标准化的事务性任务交给 AI 处理,而将需要解释力、判断力的环节保留给裁判者,实现“人机协作”的最优分工。第三,指导研发重点。由于“智”的提升可能涉及所谓的“困难问题”,研发应优先在“能”维强化数据质量、检索精度、模式识别和可解释性,在“智”维则通过法律知识图谱、因果推理模型和多模态证据分析逐步拓展推理能力。第四,风险治理与合规评估。框架可作为司法 AI 部署前的审查工具,从“能”与“智”两个维度识别潜在风险(如偏见、误判、语义歧义),并制定相应的制度与技术缓释方案。

[10] 在德雷福斯兄弟看来,从“高级新手”到“胜任者”是从“依赖外部规则与范例”走向“能够在复杂情境中自主确立判断标准”的过程。See Hubert L. Dreyfus, *supra* note[6], p.159 - 206.

(三)“智补强”之空间

按照“能—智分合”的模式或框架,目前中国司法 AI 的实践已经用五六年时间走过“能优先”阶段(阶段 1)。“行动之能”已经在程序辅助与类案推荐等场景中获得阶段性成功,尤其在争点归纳、法条匹配、文书生成等任务上展现出高度“流程化可替代性”。然而,“思考之智”仍处于严重受限状态,特别是在事实认定、规范解释与价值权衡等复杂推理环节,尚未形成具备判断能力的类人模型。这一差异表明:司法 AI 的核心挑战不在于技术是否能模仿行为,而在于能否嵌入法官的“规范性思维框架”。

当前我们进入阶段 2,即“智补强”阶段。其要求是:司法辅助型 AI 在保持高执行能力的同时,引入更完善的推理框架与知识建模,逐步增强情境感知、因果分析和价值判断能力,从而缩小“能”与“智”的差距,目前这里还存在较大的提升空间。“提升智能”的方向正在从“事务性增效”转向“可信推理”与“人机协同认知”。辅助型 AI 如果只有“能”而无“智”,就无法满足法官对可解释、可质疑、可追溯的刚性要求。结合最高人民法院《意见》的应用范围和最新实践,可对下一步提升方向作出概括和规划:第一是“智服”,达到智慧服务可感知。这涉及面向当事人和公众的可解释交互。运用司法 AI“服务多元解纷和社会治理”(《意见》第 11 条)。在司法 AI 进行诉讼风险评估、调解方案说明、执行线索查询等场景,要求 AI 用“自然语言+可视化流程图”向当事人解释“为什么这么建议”,并支持追问“如果换条件会怎样”。第二是“智管”,达到全过程风险可控。这涉及流程质量监控管理与廉洁监控管理。包括“支持电子卷宗自动分类归目、^[11]案件信息自动回填、案件繁简分流、送达地址及方式自动推荐、司法活动笔录自动生成、执行财产查控辅助、电子卷宗自动归档等智能化应用”(《意见》第 9 条),也包括用司法 AI 对庭审程序、审限、合议记录进行语义合规巡检,自动生成“异常行为报告”。比如有的法院通过全流程自动提示、智能预警和动态纠偏等,^[12]帮助院长进行事中监督,还要进行“不规范司法行为自动巡查、廉洁司法风险防控等智能化应用,提升司法管理质效,保障廉洁司法”(《意见》第 10 条)。第三是“智策”,达到智慧咨询预测的价值需求可交互可商议。涉及类案推送、裁量建议、“裁判偏离度预警、终本案件核查”(《意见》第 10 条),特别是量刑或民事和行政案件的裁量因子权重的建议,对“智”的要求更高。第四是“智推”,实现事实与规则算法推理过程的可追溯、可复盘,主要涉及推理的可信任度。“加强人工智能全流程辅助办案”(《意见》第 8 条)所涵盖的应用场景存在明显的难易程度级差,尤其是在“证据指引与审查”环节中实现可信任推理,这使得“智推”的难度更提高了一级。它必须处理多源异构证据、跨模态关联和因果推断,技术难度显著上升,且对准确率和可解释性要求极高。第五是“智识”,实现法律知识图谱与大模型的融合。把法律法规、司法解释、指导性案例、学术观点、法官伦理、审判经验等整合成可演算的法律知识图谱,与大模型参数化知识实时对齐,解决“幻觉法条”“机械教条”等问题。例如最高人民法院推动的“法信”法律基座大模型,尝试采用“知识—数据双驱动”架构,通过实现“法条—案例—学理”三元知识的对齐,确保结论的可靠性与可溯源性。^[13] 此路径的实现难度最高,不仅因其涉及动态更新的海量知识,更触及了价值共识和伦理准则嵌入的深层技术难点。而心智、意识等终极问题则有待符号主义与神经网络的未来深

[11] 参见贾宇:《论数字法院》,载《法学研究》2024 年第 4 期。

[12] 广东省政务服务和数据管理局:《经验分享:深圳在全国率先研发人工智能辅助审判,探索人工智能与社会治理深度融合应用新路径》,载广东省政务服务和数据管理局网站 2025 年 3 月 18 日, https://zfs.gd.gov.cn/xxfb/ywsd/content/post_4682295.html。

[13] 最高人民法院:《“法信法律基座大模型”研发成果新闻发布会》,载最高人民法院网站 2024 年 11 月 15 日, <https://www.court.gov.cn/zixun/xiangqing/447651.html>。

度融合,目前仍难以企及。^[14]

“辅助型”不等于“低智能”。“能—智分合”模式充分表现了我国对辅助型司法 AI 在安全可控、技术渐进、动态成长、持续探索方面的预期,如果从现在开始进行规划和行动,这一模式有望成为世界司法 AI 发展的一种有效模式。我国辅助型司法 AI 下一步重点应在“五个智”上,达到服务可感、过程可控、价值可议、推理可溯、智识可信这五个维度的持续升级状态,即“五智五可”,才能真正成为法官“用得放心、当事人看得明白”的数字助手。“能—智分合”模式应当成为未来十年司法 AI 研发与应用规划的方向、目标和任务。

三、“智”：司法 AI 的瓶颈

如前所述,司法 AI 之“能”、司法 AI 之“智”与司法具体场景任务之间是否匹配,决定了其在司法应用中的风险水平。那么,司法活动当中,存在哪些“事”是现阶段司法 AI 的“能”与“智”难以有效对应的?

司法 AI 难以胜任的任务,主要集中在具备以下五类特征的情境,这些情境共同揭示了 AI 在面对人类特有的认知、情感与价值判断时的局限: 1. 高度裁量性与价值权衡性的事务。AI 在处理涉及自由裁量权和价值冲突平衡的问题时能力有限。例如刑事量刑、未成年人保护、公益侵权赔偿等案件,法官常需在法律适用、社会效果、情理公正之间进行平衡判断。这类事务没有固定答案,需要超越逻辑演绎的“价值判断”,AI 难以胜任。2. 规范模糊与解释弹性强的事务。司法实践中大量法律条文本身具有不确定性(如“合理”“重大”“正当理由”),法官需结合上下文进行目的性或体系性解释。AI 缺乏对语义的深层理解能力,难以胜任此类需要阐释性推理的任务。3. 涉及人类情感与伦理关怀的事务。如婚姻家庭纠纷、未成年人抚养权、心理创伤赔偿等案件,裁判不仅依赖于法律条文,更依赖于情感理解、同理心和伦理关照。^[15] AI 无法理解人类情感深度和道德情境,也难以做出有温度的裁判。4. 高度个案化、无法标准化的复杂案件。如新类型案件,往往案情复杂、先例稀缺、规范不足。这类事务不具备大数据可训练的“共性”,需要人类法官发挥创造性与前瞻性判断力。5. 需要公开说理与程序正义保障的事务。司法不仅要有结论,更要有可接受的理由。^[16] AI 的“黑箱式”判断缺乏可解释性,难以满足当事人和公众对程序公正与权利保障的期望。特别是在人权、自由、身份等敏感议题上,必须由具备法律责任的法官进行充分说明。司法 AI 擅长“模式匹配”和“标准化输出”,却难以胜任“创造性解释、伦理判断、个案关照、价值冲突处理”等人类法官核心事务。这些难以被编码的能力,构成了人类法官不可替代的“司法之智”。要使 AI 在未来更好地协助裁判,就必须正视其限制并设定合理边界。

为什么司法 AI 难以胜任以上任务? 这就需要去分析司法 AI 技术所存在的两大瓶颈:

(一) 司法 AI 的因果判断瓶颈

在人类认知中,因果推断往往被视为高级智能的重要组成部分。^[17] 但从技术机理看,当下联结主义路径下的人工智能更擅长从既有数据中提取并复现统计规律,而难以把握规律背后的因果结构。这也构成了其在司法场景中难以胜任复杂推理任务的关键瓶颈。

[14] 参见魏斌:《论新一代法律智能系统的融合性道路》,载《法学论坛》2023 年第 3 期。

[15] 参见胡玉鸿:《人工智能时代法律职业“变”中的“不变”》,载《法律适用》2025 年第 10 期。

[16] 参见徐舒浩:《司法人工智能的理由模式及其功能限度》,载《法学研究》2025 年第 5 期。

[17] 参见吴小安:《机器意识与因果自我模型》,载《中国社会科学》2025 年第 9 期。

要理解这一瓶颈所产生的限制,首先需要审视人类智慧的生成。人类智慧的演化,最初表现为记忆,即把感知到的图像与事物留存。之后,随着记忆的累积,分类的需求应运而生,进而催生了类推的能力。而当信息与类别的数量增长到一定程度时,仅依赖具体的分类已不足以支撑理解与运用,于是便出现了更高层次的抽象化过程。人类通过对海量具体经验的归纳,提炼出抽象概念,并构建起一个由因果和规则组成的符号空间。这一抽象空间的建立为推理与问题解决提供了可能,也是人类智慧的核心。

“联结主义”路径下的人工智能实现了对人类智慧前半部分的模拟,即实现了记忆与分类。但是对于因果关系等更高层次的抽象思维却始终无法实现模拟。^[18]这就导致了人工智能并非同人类智慧一样通过理解因果来实现智能行为,而是通过对大规模训练数据中行为模式的归纳,总结出在特定情境下人类可能采取的反应,并据此进行行为上的仿真。它的基本逻辑,不在于是否“知道”某种行为为何正当,而在于是否在统计意义上逼近了人类行为的概率分布。例如,大语言模型之所以能够生成连贯自然的文本,并非因为它真正掌握了语言的语义规则,而是因为其在庞大的语料库中学习到了词汇与句式之间基于相关性的概率关系,从而在形式上“说得像人类”。这种情形就像一个会背乘法表,但并不理解乘法的学生,当被问及“三(乘)七”等于多少时,其可以脱口而出“二十一”。但这只是对输入、输出关系的机械记忆,一旦问题超出既有记忆范围,即便是再简单的问题其也将无法作答。大模型和深度学习系统的本质亦是如此——它们通过在海量数据中寻找相关性来映射输入与输出之间的概率关系,而并未真正理解“规则”为何存在,或“推理”应如何展开。

人工智能在因果判断上的缺陷,使得其容易陷入“过拟合”的问题,导致泛化能力不足。^[19]这也使其难以实现对法律方法的模拟,进而在处理司法问题,尤其是疑难案件时,无法举一反三,略显刻板。疑难案件之所以疑难,是因为案件事实无法被既有法律规范所直接涵摄。为此,法官要做的一个工作就是借助法律解释、漏洞填补等法律方法完善大前提,处理“事实”与“规范”之间常态化的不对称关系。^[20]这要求法官不仅应具有基于相关性的演绎能力,更能基于对历史案例和规范的理解,掌握其中的因果关系,从而将规范稳定地推导至首例案件或疑难案件之上。^[21]这是一种超越简单匹配模式的创造性的智力活动。

当下的司法 AI 由于缺乏对规范背后“因果关系”和“价值关系”的理解,其在面对疑难案件时,无法实现对法律方法的准确应用,或是在法律方法的使用上呈现出不稳定性。^[22]例如在法律解释的问题上,目前主流的通用大模型在解释方法的选取上存在随机性。不同大模型针对同一问题所采用的解释方法存在差异。即使是同一模型,在针对同一问题的多次问答中所采用的解释方法也可能有所不同。这就导致了基于法律解释的决策差异性。

此外,当下缺少因果判断能力的司法 AI 也无法在疑难案件中实现超越概率的实质性决策。一些研究发现,当大语言模型被赋予司法推理任务时,其裁判模式更为形式主义。^[23]而在司法实

[18] 参见魏斌:《论新一代法律智能系统的融合性道路》,载《法学论坛》2023年第3期。

[19] 所谓“过拟合”,即对既有训练数据中的模式形成过度依赖,从而在面对情境变化或新类型事实时缺乏稳定的泛化能力。参见魏斌:《司法人工智能融入司法改革的难题与路径》,载《现代法学》2021年第3期。

[20] 参见孙海波:《论法律的数字化与司法裁判的标准化难题》,载《行政法学研究》2025年第5期。

[21] 参见雷磊:《司法人工智能时代仍有法学方法论的运用空间吗?》,载《法学家》2025年第6期。

[22] See Andrew Coan & Harry Surden, *Artificial Intelligence and Constitutional Interpretation*, 96 University of Colorado Law Review 414 (2025).

[23] See Eric A. Posner & Shivam Saran, *Judge AI: Assessing Large Language Models in Judicial Decision-Making*, Coase-Sandor Institute for Law & Economics Research Paper No. 25-03, 2025.

践中,特别是在疑难复杂案件中,法律决策往往是需要反概率性、反相关性的。在这类案件中,法律适用常常需要突破既有经验模式,对规范间的张力进行协调,甚至对规则本身作出创造性解释。这意味着法官需要具备一种有意识地偏离经验与概率的决断能力,而这恰恰与深度学习的技术路径相悖。这也是为什么当下的司法 AI 可以实现多数案件的简单裁判,却在疑难案件中难以应对。

(二) 司法 AI 的价值判断瓶颈

更进一步,即使人工智能能够模拟人类之“智”,它也未必具有法官之“智”。一方面,司法裁判是一种规范性的价值权衡,而 AI 的底层逻辑则是描述性的统计拟合,其缺少直接进行价值判断的能力。^[24] 另一方面,法官进行价值判断时的专业思维与公众常识性思维之间也存在差异。一些学者甚至认为这种思维上的差异会导致法律人和常人头脑在生物学上的差异。法律专家在处理法律问题时更像是用规则去调控情绪,而普通人更像是用情绪去带动规则。^[25]

因此,司法 AI 不仅需要像人类一样思考,更需要像法官一样慎思。当前以“联结主义”为主的技术路径,使 AI 更像是大众“自然理性”的数字化投影。AI 训练所基于的海量数据,如互联网文本、书籍、新闻报道、社交媒体的讨论,本质上是一个充满情绪化表达的混沌集合。它忠实地记录了人类社会在“自然理性”驱动下的种种反应:直觉的判断、道德的激愤、群体的狂热与根深蒂固的刻板印象。它会通过概率分析在数据中捕捉到某种“主流倾向性”,^[26]预测哪种判决结果与语料库中绝大多数人类的情感、观点和道德偏好最为一致,并会天然地倾向于产出一个符合大众情感和直觉的输出,甚至是对输入观点进行“迎合”。^[27]这种能力使得人工智能在模拟人类日常对话、进行常识问答时表现出色,因为它的目标就是再现人类思维的普遍样貌。但这恰恰构成了它在司法场域中的致命缺陷。因为法官需要在必要时能够抵御、超越并审慎地偏离这种普遍的自然理性,以维护一种更为长远的法治价值。法官的权威与正当性并非来源于迎合当下的民意,而是来源于对法律这一“人工理性”体系的忠诚。一个真正伟大的判决,往往不是顺应舆论,而是在舆论的喧嚣中,坚定地捍卫那些超越一时一地民意的法律原则。因此,一个被设计用来学习和复制大众普遍思维模式的人工智能,其内在逻辑与法官抵御舆论、捍卫法治的天职是存在对立的。人工智能在模仿人类共识上的强大能力,反而成为其在模仿法官专业判断时最脆弱的“阿喀琉斯之踵”。

四、辅助性司法 AI“智”的三阶段路径与方案

既然司法 AI 在未来发展中始终存在“能”“智”“事”三方面不匹配的关系,我们就要设法补齐其“智”,那么实现路径是什么? 司法 AI 何以从“能”到“智”?

司法 AI 绝非披上法袍的通用人工智能,无法通过一两句简单的提示词,或一个简单的角色设定就可以实现。具有法官之智的司法 AI 需要像法官一样思考,继承法官思维的两方面特性:一是法律专业技能,即法官在法律适用过程中所依靠的专业知识体系和推理方法;二是法律职

[24] 同前注[21],雷磊文。

[25] See Takeshi Asamizuya, Hiroharu Saito, et al., *Effective Connectivity and Criminal Sentencing Decisions: Dynamic Causal Models in Laypersons and Legal Experts*, 32(19) *Cerebral Cortex* 4304 (2022).

[26] 参见李学尧:《大语言模型应用中的司法偏误与认知干预》,载《政治与法律》2025年第5期。

[27] 参见陈雅静:《警惕科研中的 AI“迎合倾向”》,载《中国社会科学报》2026年1月23日,第8版。

业伦理,即法官在裁判中所遵循的伦理规范与职业操守。前者要求司法 AI 具有超出相关性理解的识别因果、运用因果的能力,从而实现法律解释、漏洞填补等法律方法的模拟,基于可靠的因果关系精准地实现规范与事实之间的衔接;^[28]后者则要求司法 AI 能够通过价值判断来践行职业伦理,通过在具体情境中对不同价值进行权衡与选择,得以将抽象的价值具体化为裁判行动。

从人工智能技术的发展来看,当前基于深度学习的司法 AI 已展现出强大的相关性挖掘能力,能够从海量数据中识别出无数潜在的关联。^[29]例如,通过分析海量卷宗,机器可以迅速发现特定人群、特定行为与特定案件结果之间的概率关系。然而,从纯粹的概率学角度来看,任何两件事物都可能表现出一定的相关性。任何一对事物之间也都可能在概率上超越另一对事物之间的相关性。^[30]

因此,智能的真正体现不在于发现多少关联,而在于能否从海量的、真假混杂的相关关系中,筛选出那些真正具备因果逻辑的联系,并将其整合进法律推理的框架中。这就要求人工智能具有因果识别、因果判断的能力。更进一步,即便人工智能通过因果判断,在形式上掌握了法官的专业技能,也不意味着它拥有了完整的“法官之智”。因为即便确认了事物之间存在真实的因果关系,也不意味着可以将其直接作为司法决策的依据。在很多情况下,技术上所实现的所谓公平不一定能为社会价值所完全接受。^[31]司法不仅是事实判断、因果判断,更是价值判断。法律不仅是关于事实的科学,更是关于规范的实践。它处理“是什么”的问题,但最终要回答“应当是什么”的问题。因此,即使一个因果关系可以被科学证实,但它是否以及如何被纳入法律裁决,仍是一个需要价值判断的规范性问题。例如,社会学研究也许会指出,贫困、教育缺失与犯罪率之间存在因果关系。但是一个具备法律职业伦理的系统必须知道,这一因果关系绝不能成为预测某人犯罪或在裁决中产生偏见的理由。因为严格遵守公平、正义,是比探明事实因果更高级的法律要求。一个真正智慧的司法 AI,必须能够理解并内化这些核心价值,对其挖掘出的因果关系进行伦理层面的审视。它需要判断,哪些因果关系是可以在司法实践中被接受和应用的,哪些则因与法律职业伦理相悖而必须被舍弃。只有这样,司法辅助才不会因错误的价值判断而干扰到法官的决策。

当前只能识别相关关系的司法 AI,到基于“合乎价值的因果关系”进行判断的理想司法 AI,至少需要进行两方面的突破:其一是因果判断层面,其二是价值判断层面。在本文看来,司法 AI 实现这两种突破,可分为三个阶段的路径与技术方

[28] 法律推理在很大程度上表现为对因果关联的建构与审查。例如,“若无该行为,则无该损害”可被理解为一种以反事实设问为核心的因果推理方法,用以检验行为与结果之间是否具有必要联系。法律解释与漏洞填补也需要探究立法者的意图(原因)以及特定解释可能带来的社会效果(结果)。See Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*, Oxford University Press, 2009, p.392.

[29] 参见[美]伊恩·古德费洛、[加]约书亚·本吉奥、[加]亚伦·库维尔:《深度学习》,赵申剑等译,人民邮电出版社2017年版,第2—3页。

[30] 在统计学的讨论中,任何两个随时间变化的变量都可能表现出统计上的相关性,但这背后可能没有直接的因果联系,或者存在第三个“潜变量”(lurking variable)同时影响二者。See David Freedman, Robert Pisani & Roger Purves, *Statistics* (4th ed.), W. W. Norton & Company, 2007, p.150.

[31] 参见许多奇、董家杰:《论智慧税务中的法技价值对齐》,载《武汉大学学报(哲学社会科学版)》2025年第3期。

更具有意义的是通过人类的介入对基于联结主义的司法 AI 进行改造,使其输出的结果和经过法官思考后的输出结果一致,在形式上与法官的思考对齐,看起来像经过了法官的思考一样。这是在通用人工智能技术研究中正在尝试的路径,也是实现阶段 2 在“能”的基础上开始向“智”过渡的可行路径;当然,最为困难,但最具有价值的是打破“联结主义”的技术瓶颈,让司法 AI 本身具有因果判断和价值判断的能力,真正做到像法官一样思考,给出合乎价值和因果的正确输出,实现阶段 3 所期待的能智融合,目前这条路径还在探索之中。

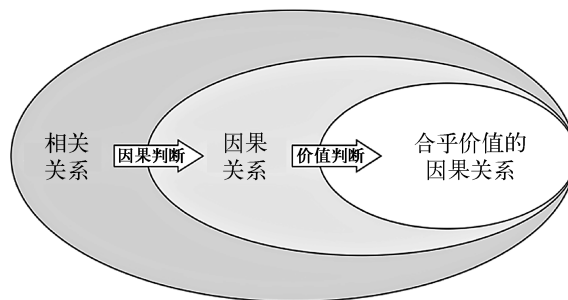


图 2 司法 AI 从“能”到“智”的路径

(一) 第一阶段：和法官一起思考

和法官一起思考,即在司法 AI 尚未实现因果识别与价值判断能力,尚处于发展“能”的阶段 1 的背景下,将追求人工智能实现、超越法官之智的目标,暂时由人机协同系统替代。确保人与机器融合形成的“增强型”司法系统,能够在决策的质量与效率上超越普通法官的能力。这种路径的根本特征在于:人工智能不再被视作“替代者”,而是以“共思考”的角色出现,即通过具有可解释性的输出与法官的思维形成互动,使司法 AI 与法官一同思考。^[32] 人工智能为法官提供计算支持,而法官则出于未决个案的特殊性,决定是否采纳,并以专业知识与价值判断对人工智能的推理进行校正与引导。^[33] 在现阶段技术条件尚不成熟的背景下,应确保司法智能化既能发挥效率优势,又不偏离司法裁判所需的因果和价值判断。

首先,与机器合作的前提是了解机器。只有当人工智能打破“黑箱”,通过可解释性机制将其推理路径、数据基础和逻辑假设透明化时,法官才能在关键节点进行有效的质疑、修正与确认,为其注入人类独有的因果判断和价值权衡。因此,实现法律人与司法 AI 合作的前提在于“可解释性”^[34]。

其次,在可解释性基础上,人机协同的司法模式的特征是“思维共生”。这意味着司法 AI 并非意图替代法官,而是通过解释引导法官进入更高效、更全面的思维过程,帮助法官更全面地把握案件事实与推理方向,为法官提供“可供思考的素材”^[35]。而法官则赋予这些素材以“规范意义与价值正当性”,在充分了解机器所依托的相关性后,基于其专业素养对这些素材加以确认和筛选。例如,当人工智能通过数据挖掘得出某一行为模式与犯罪结果高度相关时,其责任并非推动法官采纳,而是通过解释性机制提示该相关性可能隐藏某种因果逻辑。此时,法官必须结合法律方法展开追问,探究其背后真正的社会根源与法律意义。一是对人工智能提示的相关性进行因果性验证,剥离表象的统计关联,追问更深层次的制度性、结构性原因,从而避免将伪因果纳入裁判;二是

^[32] “人机协同”或“人在回路”(Human-in-the-Loop)是当前阶段发展负责任人工智能的共识路径。它承认纯粹自动化系统的局限性,强调人类的监督、引导和最终决策权。在司法等高风险领域,这种模式被认为是确保 AI 应用安全、合乎伦理的必要保障。See Saleema Amershi, Maya Cakmak, W. Bradley Knox & Todd Kulesza, *Power to the People: The Role of Humans in Interactive Machine Learning*, 35(4) AI Magazine 105 (2014).

^[33] 参见周翔:《刑事辩护变革的契机:人机协作司法模式》,载《交大法学》2025 年第 6 期。

^[34] See Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 Columbia Law Review 1829 (2019).

^[35] See Burkhard Schafer & Zhaoning Li, *Explainable AI and Law: An Evidential Survey*, 32 AI and Law 1 (2024).

在价值层面对人工智能提示的因果链条进行合法性与正当性检验,思考哪些因果解释是可以在司法领域被接受的,哪些则必须被排除在外;三是考虑因果解释背后的社会根源与法律意义。例如,如果人工智能揭示“教育机会不足与青少年犯罪高度相关”,法官不能简单地把教育缺失当作个体的风险标志,而应意识到这一因果关系所揭示的是制度性问题。在这种情况下,说理需要体现出对社会根源的关注,并通过法律语言将其转化为对公平与正义的回应。

最后,基于可解释性的人机协作,可以为司法 AI 的未来发展提供过渡性平台。现阶段,人工智能无法独立完成因果识别与价值判断,而由人辅助思考的模式恰恰为未来的演进积累了经验与资源。每一次人机协同中的解释、质疑与修正,都是对人工智能的训练。^[36] 随着时间的推移,人工智能将逐渐在逻辑结构与推理模式上接近法官的思维方式,进而在未来实现更高水平的自主智能。

(二) 第二阶段:与法官的思考对齐

与法官的思考对齐,即基于法律专业技能与职业伦理实现司法 AI 的价值对齐。其要义在于使人工智能的输出在形式上尽可能贴近法官的推理逻辑,使其结论逐渐在表面上与法官的思维路径趋同,从而在外观上具备“法官之智”^[37]。这要求司法 AI 应当处于一个能够持续审查和校准的机制之中,使其在因果与价值层面能够与法官判断保持一致。^[38] 这种方式规避了人工智能所不可能实现的价值和因果判断,转而通过“训导”的方式使其所具有的相关性判断与人类所具有的因果判断和价值判断对齐。^[39]

机器的相关性关系与因果、价值关系对齐的起点在于提升可解释性。对于不熟练掌握计算机知识的法律人而言,一个不透明的“黑箱”模型既无法被观察和审计,也无法进行有效的干预。^[40] 可解释性的功能在于将模型内部的运算过程,转译为人类可理解的表达,为法律人的介入提供接口,从而对模型输出所依赖的关联关系进行连续的审查与校准。^[41] 这种技术性的呈现,使得模型的统计学“论据”变得透明化,将原本封闭的计算过程转化为一个开放的、可供分析和干预的对象,这是后续所有对齐工作的起点。

基于可解释性,法律人对司法 AI 的审查存在两个方面:其一是基于法律专业技能的因果性审查,其目标是使模型的推理路径符合事实认定和事实与规范连接的技能要求。法律人运用其专业知识对基于算法解释揭示的关联因子列表,进行因果审查。首先,排除统计上显著但逻辑上不成

^[36] See Pascal Schramowski, et al., *Making Deep Neural Networks Right for the Right Scientific Reasons by Interacting with Their Explanations*, 2 *Nature Machine Intelligence* 476 (2020).

^[37] 例如大语言模型通过在海量文本数据上进行训练,学会了生成在形式上(语法、风格、逻辑结构)与人类语言高度相似的文本,但它们并不“理解”这些文本的深层语义或其与现实世界的因果、价值关联。See Yoav Goldberg, *A Primer on Neural Network Models for Natural Language Processing*, 57 *Journal of Artificial Intelligence Research* 345 (2016).

^[38] 一些观点认为,既然无法让 AI 内生性地理解价值,就只能通过外部的、持续的校准机制,强制其输出行为与人类的价值偏好保持一致。这也构成了 RLHF 和 Constitutional AI 等技术的基本思想。See Christoph Winter, Nicholas Hollman & David Manheim, *Value Alignment for Advanced Artificial Judicial Intelligence*, 60 *American Philosophical Quarterly* 187 (2023).

^[39] 参见王沛然:《从控制走向训导:通用人工智能的“直觉”与治理路径》,载《东方法学》2023 年第 6 期。

^[40] See Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, “*Why Should I Trust You?*”: *Explaining the Predictions of Any Classifier*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, p.1135 - 1144.

^[41] 参见周翔:《算法可解释性:一个技术概念的规范研究价值》,载《比较法研究》2023 年第 3 期。

立的虚假关联,这类关联通常不具备任何现实世界的因果基础。其次,识别并处理由未观察到的混杂变量所导致的误导性关联,这通常需要引入外部知识、专业知识来判断该关联是否掩盖了更深层的、真实的因果结构。最后,确认并保留那些能够与证据规则、逻辑规则和经验法则相印证,且能够被整合进一个连贯的因果叙事中的关联因子。此过程的实质,是将人类的因果推理能力作为一种外部过滤器,作用于机器生成的、未经筛选的关联性集合之上,从而确保模型在独自构建其“事实”基础时,其所依赖的特征在因果逻辑上是有效的。

然而,通过因果性审查的司法 AI 并不意味着在司法场景下必然是可接受的。第二层干预是价值层面的对齐,旨在使模型的决策符合法律职业伦理的约束,在价值多元、冲突与碎片化的现实条件下实现价值排序和取舍。^[42]这不仅要求事实认定层面的准确性,更要求决策过程和结果的正当性。在这一阶段,法律人应依据内化的职业伦理规范和法感,对已经通过因果审查的因子进行二次过滤。某些因子即便在因果上是成立的,但若其与基本的法律原则相冲突,则必须禁用。例如在量刑模型中,即便数据表明个体的某些社会经济背景特征与其再犯率存在因果联系,但基于价值判断,这些特征亦不能被用作判断犯罪和加重处罚的依据。

在技术上,当前实现相关性与因果和价值层面的对齐可以通过两种路径进行:其一是基于人类反馈的强化学习(RLHF)路径。^[43]其基本逻辑是引入人类对人工智能判断的评价,将人工智能的行为与人类判断进行比较,通过不断地正向奖励与负向惩罚,引导其优化输出。通过持续地反馈训练,模型的目标函数逐渐向人类专家的因果偏好和价值排序收敛。^[44]其二则是基于“宪法性 AI”(constitutional AI)的路径,即借助人工智能自我监督的方法实现人工智能的价值对齐。例如通过一个从属的 AI 模型,评估主模型的输出是否遵循了特定的“宪法性”原则(即一套事先确定的原则或规则),进而用于优化主模型。通俗地讲,既然司法 AI 无法进行价值判断,那么就提前将司法场景中可能遇到的价值判断以答案的形式输入人工智能。^[45]

(三) 第三阶段:像法官一样思考

值得注意的是,用相关性判断替代因果判断和价值判断只能实现看起来像人类一样思考,却无法真正实现像人类一样思考。司法 AI 要想走向“能智融合”的阶段 3,就不能只筛选相关性,而应在未来由司法 AI 自主实现对因果判断和价值判断的模拟。^[46]

2011 年图灵奖得主朱迪亚·珀尔(Judea Pearl)认为,当前基于联结主义的学习方式是一种低层次的认知,亟须通过主动干预和反事实推理来去伪存真。^[47]为此,他将因果推理概括为从低到高的三个认知层次。第一层是通过“观察”(Seeing),实现对“关联”的发现,例如观察“罪犯身份证号”与“犯罪率”之间的关联性。第二层是“干预”(Doing),即人为改变其中的某个条件,观察结果

[42] 参见矣晓沅、谢幸:《大模型道德价值观对齐问题剖析》,载《计算机研究与发展》2023 年第 9 期。

[43] 参见韩旭至:《大模型价值对齐的法治进路》,载《中国法律评论》2025 年第 1 期。

[44] See Long Ouyang, Jeff Wu, Xu Jiang et al., *Training Language Models to Follow Instructions with Human Feedback*, in *Advances in Neural Information Processing Systems*, Curran Associates, 2022, p.27730.

[45] 在已有的实践中,Claude 等模型证明了宪法性 AI 方法的有效性,即帮助其减少有害的、歧视性的输出,并对使用者的“对抗性输入”作出更恰当的非简单采取回避策略。See Yuntao Bai, Andy Jones, Kamal Ndousse, et al., *Constitutional AI: Harmlessness from AI Feedback*, arXiv: 2212.08073 (2022).

[46] See Judea Pearl & Dana Mackenzie, *Judea Pearl, AI, and Causality: What Role Do Statisticians Play?*, *Amstat News*, September 2023, p.6 - 9.

[47] 珀尔认为,要从数据中获得因果知识,仅靠被动观察(seeing)是不够的,必须引入“do-算子”(do-calculus)来模拟主动干预(doing),并最终能够进行反事实推理(counterfactual reasoning)。参见杨强、范力欣、朱军等:《可解释人工智能导论》,电子工业出版社 2022 年版,第 20 页。

是否会随之变化。比如,强行规定所有身份证号都以偶数结尾,犯罪率真的会下降吗?这种设问有助于识别哪些因素是真正的因果条件,哪些只是伪关联。第三层是“反事实”(Counterfactuals),即在既定事实之外进行想象。例如,假设一个已经被判刑的人并非出生在经济条件落后的环境中,而是在资源更丰富的环境中成长,他是否还会走向犯罪?^[48]

司法裁判的精髓正是在因果阶梯的最高处运作,一个真正具备法官之“智”的司法 AI 也必然是一个能够在因果阶梯上自由攀爬的模型。为此,司法 AI 在技术上必须满足三方面的要求:其一是具备自我解释的能力,能够清楚标明自身结论所依赖的相关性依据,并识别其中可能的脆弱性,从而避免成为一个连自己都不理解的“黑箱”。其二是具备自我干预的能力,能够在推理过程中主动改变某些条件,并据此检验结果的变动方向和幅度,以此来验证因果链条的稳固性。其三是具备反事实推演的能力,能够在既定事实之外构建假设情境,并检视不同假设下的潜在结果,从而判断某一结论在司法实践中的适用边界与合理性。^[49]

然而,当技术层面的要求得以实现后,一个更严峻的问题在于:阶梯的顶端通向何方?司法 AI 的目标——那个由法律专业技能和职业伦理构成的“法官之智”的顶峰——其内部的知识结构、推理路径和价值坐标,是否已经被我们清晰地描绘出来了?答案恐怕不容乐观。一方面,长期以来,法官的智慧在很大程度上是一种“默会知识”,它根植于经验、直觉和难以言传的职业技艺之中。^[50]另一方面,一些法学问题还存在争议。这就导致即便 AI 能够进行因果和价值判断,其也可能因掌握了不同的法学流派观点而产生不同的判决。

因此,通往“能智融合”的终极路径,必然是一场技术探索与法学自我求索的双向奔赴。在计算机科学家努力让机器向上攀爬的同时,法律共同体必须承担起向下挖掘、向内反思的使命:将那些暗默的、不自觉的裁判知识“显性化”“结构化”和“知识图谱化”。^[51]这使得我们需要重新回答一系列根本性的问题,例如法律因果的构成要件与排除规则究竟是什么?在面对具体的价值冲突时(例如,效率与公平,自由与秩序),法律职业伦理所要求的权衡逻辑和优先次序如何因场景而确定?法官在事实叙事中,是如何将证据碎片整合为一个具有内在因果一致性的法律事实的?这些问题,如果法律人自己都无法给出一个清晰、自洽的答案,又怎能期望人工智能通过代码实现?

五、结 语

本文认为,在最高人民法院确立“辅助审判原则”之后,提升司法 AI“智”的维度依然具有必要性与可行性。核心观点是强调:通过“能—智分合”演进模式,提升辅助型 AI 的智能,这契合中国司法 AI 发展的实际特征和政策预期,有望成为中国司法 AI 的可行的成长模式。

在过去五六年中国司法 AI 主要处于“能优先”的阶段,侧重流程化与标准化任务的效率提升;未来十年则应进入“智补强”阶段,在因果推理、情境理解、价值判断与知识图谱整合方面取得突破,以回应法官对可解释性、可追溯性与正当性的需求。展望未来,司法辅助型 AI 的发展

[48] See Judea Pearl & Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect*, Basic Books, 2018, p.28.

[49] 参见吴小安、俞沁元:《大语言模型与因果之梯》,载《自然辩证法通讯》2025年第8期。

[50] See Michael Polanyi, *The Tacit Dimension*, University of Chicago Press, 2009, p.4.

[51] 法律知识图谱化是指将法律文本中蕴含的复杂概念、规则、事实和关系,用机器可读的结构化语言进行表征,从而为更高级的法律推理提供基础。

愿景是在“五个智辅”方面具备服务可感、过程可控、价值可议、推理可溯、知识可信等“五可”特征。

在“能—智分合”演进模式下,中国司法 AI 在“智”维度的未来发展可细分为三个阶段性路径:第一步是“与法官一起思考”,通过可解释机制支撑人机协作,避免 AI 越权。第二步是“与法官思考对齐”,借助反馈学习,使 AI 的结论在形式上贴近法官逻辑。这两步可能需要十年甚至更长的时间。第三步是“像法官一样思考”,即在因果推理、反事实演绎和价值判断上实现真正突破,达到“能智融合”的高级状态。这需要更长的时间。因为它需跨越三个瓶颈:一是突破“相关性难以跃迁为因果性”的技术困境;二是克服 AI 缺乏法律职业伦理与价值判断的不足;三是解决大众常识与法律专业思维之间的鸿沟。

Abstract Even after judicial artificial intelligence has been explicitly positioned as “assistive”, the future trajectory of judicial AI in China should still aim to enhance the “intelligence” of assistive judicial AI. Building on the two-dimensionally separable relationship between AI’s “capability” and “intelligence”, and drawing on the Dreyfus model, this article proposes a “capability – intelligence separation-and-integration” model for the development of China’s assistive judicial AI. At present, further improvements in judicial AI’s “intelligence” face technical bottlenecks in causal judgment and value judgment, but this does not mean that viable technical pathways are absent. The enhancement of “intelligence” in China’s assistive judicial AI can be advanced in stages through three technical schemes. The first is “thinking alongside judges”, using explainability as a key lever to achieve a formal breakthrough in practical usability. The second is “aligning with judges’ reasoning”, promoting value alignment under the constraints of legal expertise and professional ethics. The third is “thinking like judges”, ascending the ladder of causal and value judgment so as — within the assistive mandate — to strengthen inferential support for complex discretionary adjudication.

Keywords Judicial AI, Assistive Judicial AI, Dreyfus Model, Explainability, Value Alignment

(责任编辑:徐舒浩)