

# 刑事智能预判与法官裁判的认知冲突及其协调

韩振文<sup>1</sup>, 周芷妍<sup>2</sup>

(1. 浙江工商大学 法学院, 浙江 杭州 310018;

2. 华东政法大学 刑事法学院, 上海 201620)

**摘要:** 伴随着生成式人工智能的飞速迭代升级, 人工智能与刑事司法的深度融合已然成为显性趋势。但不可忽视的是, 当前刑事智能预判的运行逻辑与法官裁判的认知过程仍然存在显著差异, 体现为智能预判无法推理因果关系、难以认定案件事实且缺乏实践理性。技术上的局限决定了智能系统作出的预判结论可能与法官的裁判思维在认知层面产生冲突, 其一方面通过模式识别偏差干扰法官的决策, 另一方面使法官因普遍存在的认知纠缠倾向而陷入对智能预判结论迷信盲从与怀疑抵触的思维反复。因此, 需要遵循技术性正当程序规制理念, 基于刑事案件复杂度动态限定调整智能预判的应用边界, 优化预判结论的可解释性, 并构建法官分类分级参考机制, 以尽可能调和刑事智能预判与法官裁判的认知冲突。

**关键词:** 刑事智能预判; 法官裁判; 认知冲突; 技术性正当程序

**中图分类号:** D914

**文献标识码:** A

**文章编号:** 1672-1217 (2026) 01-0117-12

**收稿日期:** 2025-11-03

**基金项目:** 2024 年度国家社会科学基金一般项目(24BFX036): 智能预判的技术性正当程序控制研究。

**作者简介:** 1. 韩振文 (1987-), 男, 山东滨州人, 浙江工商大学法学院副教授, 法学博士;

2. 周芷妍 (2001-), 女, 浙江温州人, 华东政法大学刑事法学院硕士研究生。

## 一、问题的提出

近年来, 在我国人工智能与司法融合发展、法官普遍承受案多人少的压力以及司法能力危机等多重因素交互作用的背景下, 刑事智能预判系统已成为审判能力现代化的需要。智能预判, 亦称算法裁判, 是依托人工智能对裁判进行自动预测生成并精准推送的辅助办案技术。<sup>①</sup> 预判结果理论上包括针对具体案件的裁判结论, 以及用自然语言为其提供相应的释法说理。智能预判属于智慧法院建设中算法自动化决策的核心体现, 在类案检索、裁判偏离度预警、电子卷宗随案同步生成等技术加持下, 其在刑事司法领域将逐步得到广泛适用。

各地法院在实践中已陆续研发各具特色的刑事智能预判系统。例如, 上海“206 系统”尝试在盗窃、电信网络诈骗和非法吸收公众存款等常见罪名中实现证据标准指引、量刑建议参考、社会危险性审查等功能;<sup>②</sup> 浙江法院“凤凰智审”开发危险驾驶罪智能化审判模块以及诈骗罪要素式智能审判工作, 探索案情要素提取、量刑计算、裁判文书自动生成等功能。这些系统在简单、重复性高的刑事案件中展现出效率提升的优势, 为法官提供了重要的参考工具, 但同时也暴露出在案件证据庞杂、牵涉面广、被害人众多等情境下表现不佳, 存在数据偏见和算法缺陷等风险。因此, 应当明确刑事智能预判系统的价值在于增强而非取代法官的裁判能力, 其未来较长时间内都将定位于法官辅助性工具, 所生成的结论只能作为参考意见而非裁判的直接依据。

2024 年 11 月, 最高人民法院发布“法信法律基座大模型”, 其技术创新主要体现在探索如何遵循司法规律, 将我国司法审判业务需求与大模型技术相结合, 让大模型“先通后专, 通专结合”, 最终实现对法律

<sup>①</sup> 韩振文:《智能预判应用风险的技术性正当程序控制》,《中国社会科学报》2024 年 7 月 4 日, 第 A04 版。

<sup>②</sup> 严剑漪:《揭秘“206”: 法院未来的人工智能图景——上海刑事案件智能辅助办案系统 164 天研发实录》,《人民法治》2018 年第 2 期。

业务的可靠支撑。<sup>①</sup>理想情况下,作出裁判预测的算法模型不仅需要精通法律知识、了解行业惯例、掌握社会常识,更应当具备人类的逻辑推理能力与价值判断能力,这样才能在司法实践中发挥实质性的功用。然而,当前的刑事智能预判系统或许尚难以达成这一目标。已有的系统都需要依靠开发者来把握核心问题,选定数据范围,并对算法模型反复试验调校;同时,其很大程度上依赖人工将大量非结构化数据(文本、语音、视频)进行结构化。<sup>②</sup>换言之,在刑事智能预判系统的设计和运行中,“人工”所占的比重仍远大于“智能”。显然,司法人工智能还未能达到人们科幻化的期待。

刑事司法裁判过程中,智能预判会对刑事证明的认知结构与权力结构产生显著影响。<sup>③</sup>人工智能通过大数据分析和算法模型生成的证据分析、量刑建议及人身危险性评估,可能在不同程度上影响法官的决策方式,这种影响既可能表现为法官模式识别偏差的加剧,对案件产生错误认识;也可能体现为法官对智能预判因其工具理性存在盲目依从倾向,同时又因其难以被理解和解释的特性而怀疑不安,对应当如何看待和处理预测结果感到无所适从。更为重要的是,这些影响不仅停留在法官认知层面,还会侵蚀传统的程序正义理论,干扰司法公正。

鉴于此,本文运用认知心理学的分析工具,着眼法官独立决策向智审辅助人机协同决策转变的现实情况,深入探析刑事智能预判形成过程与人类认知思维的差异,充分检视刑事智能预判与法官裁判耦合过程中在认知层面可能产生的冲突及根源。在此基础上结合现行诉讼程序法理和有效规范,以基于主观程序正义的技术性正当程序为轴心,提出刑事智能预判与法官裁判的认知冲突调和方案,以期实现二者之间的协调发展。

## 二、刑事智能预判形成过程相较法官裁判认知过程存在的缺陷

如今的刑事智能预判系统所取得的成就无疑是巨大的,但其预判结论的形成过程与人类法官的认知机制始终存在根本差异。法官的裁判决策表现出鲜明的实践理性品格,而智能预判则依赖于技术理性的刻画与表达。尽管智能预判能够减轻法官认知负担,发现案件中的潜在规律,但仍反映出诸多技术局限。根据不同的发展阶段,人工智能可以被分为强人工智能和弱人工智能。前者有着感知和自我意识,能够自发学习知识作出决策、处理问题;但其不仅在技术实现上面临着巨大挑战,在伦理上也存在很多争议。<sup>④</sup>与此相对的是后者虽然无法自发感知与自主学习,但可通过算法分析数据,实现特定推理与预测。刑事智能预判系统作为典型的弱人工智能,依赖统计建模进行模式识别,其作出的预判结论基于数据相关性而非因果推理,面对疑难复杂案件无法判断事实真相,更难以基于实践理性作出裁判决策,这些差异必然引致其预判结论的合理性与公正性受到质疑,对传统的程序正义理论造成冲击。

### (一) 无法推理因果

因果推理是人类认知的重要特征,源于复杂的认知结构,包括直觉推理、反事实思维以及对特定语境的理解。<sup>⑤</sup>人类法官的裁判会综合考虑案件中的事实逻辑、法律关系、社会背景,进行因果推理。法官能判断“X发生,则Y成立”,并能解释“假如没有X,Y是否仍然成立”。因果关系能够揭示事物之间关系的本质,让人们避免落入数据相关性的陷阱。

在强人工智能尚未实现的今天,现有的刑事智能预判系统主要依赖于统计学习和数据模式匹配,无论其采用专家系统还是机器学习技术,本质上都只能基于数据特征相关性作出归纳或演绎,并不具备真正的因果推理能力。<sup>⑥</sup>具体表现为,智能预判系统通过对与案情有关的海量历史信息进行深度挖掘、拟合分析

①张昊:《法律基座大模型将带来哪些“智体验”》,《法治日报》2024年12月19日,第6版。

②魏斌:《法律大语言模型的司法应用及其规范》,《东方法学》2024年第5期。

③沈磊:《人工智能对刑事证明结构的影响及其制度因应》,《法学研究》2025年第6期。

④莫宏伟:《强人工智能与弱人工智能的伦理问题思考》,《科学与社会》2018年第1期。

⑤梅剑华:《人工智能与因果推断——兼论奇点问题》,《哲学研究》2019年第6期。

⑥丁梦远、兰旭、光彭茹、郑南宁:《机器推理的进展与展望》,《模式识别与人工智能》2021年第1期。



寻找相关性,发现人类司法裁判行为的内在规律,并将之应用于对未来类案裁判的预测,<sup>①</sup>却并不理解有关证据之间以及证据与裁判结论间的因果关系。仅关注数据的相关关系而不去探究事物背后的因果逻辑,往往会得到令人不解的结论。问题的核心在于,智能预判能够发现“X发生时,Y往往可能发生”,但无法判断“X是否是导致Y的真正原因”,这意味着其生成的内容可能严重偏离现实情况。

这一缺陷直接影响了预判结论的公正性与准确性,可能导致刻板印象的强化以及预判结论的偏差。尤其在疑难复杂案件中存在诸多要素叠加,需要对其拆解分析并反复斟酌,难以通过相关性的对应关系轻易作出结论。在量刑建议中,这一缺陷尤为突出,例如系统可能仅因“累犯”标签就推送更高量刑区间,却无法解释其是否必然导致以及为何导致刑罚加重,其逻辑链条缺乏因果证成。而在人身危险性评估中,刑事智能预判系统则可能错误地在经济条件、教育程度等要素与再犯风险之间赋予直接因果关系,这种相关性替代因果性的逻辑会固化偏见,带来制度性歧视。此外,由于智能预判系统缺乏反事实推理能力,其无法回答在前提条件改变的情况下,结果是否也会变化。在前面的例子中,当智能预判遇到“如果被告人的经济条件与社会背景良好,其是否仍会犯罪?”此类问题时,其无法评估假设情况对裁判结论的影响,而这往往是法官在决策中需要综合考量的重要因素。

此外,刑事智能预判系统不仅需要作出裁判结果预测,还需要合理地解释该结论如何证成,使其具备法律适用上的可解释性和正当性。预测可以以相关关系为基础,而解释和推断的基础必然是因果关系。<sup>②</sup>这意味着现有的智能预判系统即使能够较为准确地作出裁判结果预测,也难以对这种决策的原因进行理解和说明,在生成裁判文书的说理部分时只能进行机械套用。这种局限性不仅影响预判的公正性,可能面临伦理和法理上的质疑,也使法官难以对智能预判结论的合理性产生信任。

## (二) 难以认定事实

法官在认定案件事实过程中,需要对在案证据进行综合审查分析,这需要经过文本内容理解、证据规则检验、整体逻辑串联等多个环节。对此,当前的刑事智能预判系统均难以完全掌握,由此导致了其核心缺陷,即无法做到像人类法官那样感知证据、理解案件事实。<sup>③</sup>这一缺陷涉及通用人工智能领域的基础问题:如何让人工智能像人类一样理解复杂的故事。叙事和理解叙事是人类智能的重要功能,也是法律工作的核心任务。从某种角度看,法律程序正是在重新构建一个过去发生的故事,故事的基础是事实,事实的基础是证据。<sup>④</sup>然而,证据的真实性、事实的可靠性,都需要法官根据自身的经验法则和逻辑法则进行判断,形成心证确信。从某种意义上说,判断一个故事的真假,才是法律决策的真正难点。

以于欢案为例,该案核心争议在于被告人是否构成正当防卫。这一判断依赖于对案发环境(母亲遭受凌辱、多人围堵)、被害人行为(是否持续侵害)、被告人主观意图(防卫还是报复)的综合认定。在事实认定环节,这些要素主要是通过对监控视频、证人证言、被告人供述等证据材料的分析而确立的。为实现这一目的,刑事智能预判系统需要突破重重困难:识别在案证据是否存在篡改或伪造的情形,在模态众多的基础材料(文本、图像、音频、视频)中摘录出关键信息,<sup>⑤</sup>理解每一个信息对于判断案件事实的意义,最后将这些材料综合形成对事实的整体性理解。

现实情况却是,刑事智能预判系统在面对深度伪造(Deepfake)的音视频或修改过的电子证据时,难以有效鉴别其真伪,仍需依赖人工复核;<sup>⑥</sup>在多模态信息的处理过程中,系统在整合提取时可能遗漏关键信息;法律语言的模糊性和法律场景的特殊性使得系统难以准确理解语境以及言语表达中的细微差异。<sup>⑦</sup>尤其是证人证言和书面证据中往往包含情绪和语义歧义,而刑事智能预判系统受制于有限的自然语言处理能力,只能进行字面分析,无法识别微妙的言外之意,难以准确判断文本内涵的真实性。于欢案的关键在

①雷磊:《司法人工智能能否实现司法公正?》,《政法论丛》2022年第4期。

②刘庄、卢圣华:《机器能取代法官吗?:人工智能、数据科学与法律》,北京:北京大学出版社,2025年,第201页。

③谢登科、周鸿飞:《司法人工智能在证据审查中的功能定位与风险规制》,《吉林大学社会科学学报》2024年第3期。

④张保生:《事实、证据与事实认定》,《中国社会科学》2017年第8期。

⑤刘庄:《幻象与本相:法律人工智能及其他》,《中国法律评论》2024年第2期。

⑥王禄生:《论“深度伪造”智能技术的一体化规制》,《东方法学》2019年第6期。

⑦周尚君、伍茜:《人工智能司法决策的可能与限度》,《华东政法大学学报》2019年第1期。

于“辱母”“围堵催债”等场景的语境理解与情感张力,类似的复杂语境普遍存在于正当防卫案件中,但智能预判系统的语言处理能力无法识别被害人过错、具体场景、行为手段、加害与被害的心理情感等影响正当防卫是否成立的实质性判断要素,将导致定罪基础严重偏差。此外,前文所述的智能预判系统在因果推理能力上的匮乏,决定了其对案件的理解仅停留在表层数据统计,这显然不足以支撑其完成对案件事实的认定。

刑事智能预判系统在事实认定方面的局限,或许还可以从证据法学研究中窥见端倪。西方学者们曾尝试通过数学模型来量化证明力,然而,证明力评估判断的本质并非纯粹的数学计算,而是一种高度依赖裁判者理性和经验的主观判断。即便是“新证据学派”所倡导的贝叶斯方法,也难以克服如先验概率赋值的主观性等问题,由此引发了贝叶斯主义与培根主义的长期争论。<sup>①</sup>这一争论至今仍在西方法学界持续,充分说明了证明力判断无法单纯依赖数据和算法量化,更不可能由智能预判独立承担这一重任。因此,面对复杂多样且不断进化的证据形态,刑事智能预判系统显然没有能力冲破庞杂证据的重重迷雾,从中准确地认定案件事实。

### (三) 缺乏实践理性

诚如霍姆斯所言,“决策的真正基础是对政策与社会利益的考量,假设解决方案仅仅通过无人争议的逻辑或者一般法律命题就能得到,那将是徒劳的。”在刑事领域,法律决策除了涉及复杂的因果推理和事实判断,更涉及道德原则、正义观念以及社会价值的维护,同时刑法中存在许多酌定情节,这些情节的幅度和范围该如何确定,都需要依据案件本身的具体情况,由法官来进行衡量和判断。<sup>②</sup>在没有明确法律规定或先例的疑难案件中,法官需要通过实践理性填补规则的漏洞,这种自由裁量的依据正是法官自身的社会经验,以及对法律背后价值判断的认识。与之相对的,智能预判的优势在于其工具理性品格,这种模型无法兼容对价值理性的追问,<sup>③</sup>无法对概念进行认知性理解,更无从奢望其能像人类一样将感觉经验与意识先验相结合,在裁判中对各种价值作出基于社会现实环境的衡量。

尽管司法实践中正努力推动智能预判系统实现先“通才”后“专才”,类似于人类知识学习过程,即先接受广泛的基础教育,再接受法律专业培训,最后获取特定任务的实践经验。<sup>④</sup>但对刑事个案要素作出道德判断和价值考量,并以此弥补法律空白,明显超出了机器学习的能力边界。尽管人们已在价值衡量的相关问题上反复讨论,但对于不同社会价值的属性特征仍没有形成清晰的充分认识,<sup>⑤</sup>由此便难以实现价值的数据化。同样代入到于欢案中,于欢无前科、无持续暴力倾向,行为属于被害人辱母造成的道德冲击及群体围堵所形成的心理压力等特定情境下的防卫,不具有普遍危险性,但其“致一人死亡,两人重伤、一人轻伤”又属于危害后果严重。在类似“严重暴力犯罪”的认定上,社会价值权衡能力的缺乏,使得智能预判系统无法综合考量暴力程度、危害后果以及主观恶性等因素,作出符合人们对于正义普遍期待的预判结论。

实践理性的基础是大量的生活经验与司法经验,在刑事审判的过程中,法官需要调动的经验数量是无穷的,包含了道德教化、群体互动、传统文化等等。<sup>⑥</sup>在这些经验中,事实和价值做不到泾渭分明的区分,<sup>⑦</sup>而且其内涵和外延均会随着时间的改变以及社会的推演而不断变化,所以任何社会都不可能将能够运用的经验做完全的归纳和列举。虽然智能预判系统能够通过大数据或机器学习发现一些统计学意义上的“经验”,但其所能获得的数据相对于整个社会来说毕竟是有限的,无法涵盖每个领域。<sup>⑧</sup>因此,智能预判无法

---

①[美]特伦斯·安德森,戴维·舒姆,[英]威廉·特文宁:《证据分析》,张保生等译,北京:中国人民大学出版社,2012年,第305—307页。

②王申:《法官的实践性知识与智慧导向》,《法制与社会发展》2011年第6期。

③胡昌明、张吟:《中国司法人工智能面临的挑战及应对》,《学习论坛》2025年第2期。

④余茂玉、林志农、张承兵、石鹏:《法律基座大模型的建构路径与应用展望》,《数字法治》2025年第1期。

⑤彭中礼:《司法裁判过程中的人工智能应用研究》,北京:法律出版社,2024年,第132页。

⑥柯华庆:《法律变革的逻辑——社科法学与教义法学的较量》,《桂海论丛》2014年第4期。

⑦[日]松尾丰:《人工智能狂潮——机器人会超越人类吗?》,赵函宏、高华彬译,北京:机械工业出版社,2016年,第66—67页。

⑧纵博:《人工智能在刑事证据判断中的运用问题探析》,《法律科学(西北政法大学学报)》2019年第1期。

将人类能够运用于司法的经验全部植入或模拟，于是也难以具备基于个案情况进行合理决策的实践理性。

同时，由于经验的可靠性难以判断，法官必须依靠实践理性的指导才能恰当运用经验。只有合理的经验才能用于司法决策，产生合理的裁判结果。实践理性意味着法官可以自我反思、自我审视，通过道德思考来突破先前的偏见和错误。<sup>①</sup>然而，智能预判系统本质上是模式识别工具，无法突破数据局限，进行道德反思，意识到自身数据库中的经验素材可能存在的问题，并主动调整。由于社会经验整体是包含着科学事实、印象性信念、刻板印象的复杂混合体，价值判断可能会在其中伪装成经验命题，其能否在刑事审判中被接受有着特定的标准。面对复杂的价值判断问题，实践理性的匮乏使得智能预判系统只能机械裁判，难以作出符合个案正义的判断。

### 三、刑事智能预判与法官裁判的认知冲突

学界关于智能预判风险的讨论正如火如荼，对其输出结果的担忧多集中在“算法黑箱”减损可解释性、“算法歧视”巩固历史偏见、编造案例扰乱司法公正等方面，<sup>②</sup>却基本忽视了智能预判作为法官裁判决策的一种参考性意见，在庭审中意见输出表现出“法官在前、机器在后”而在文书初稿生成阶段体现出“算法前置、法官后撤”的样态，其无论何阶段给出的意见本身合理与否，都会参与到法官的认知过程中产生认知共振，对法官的决策产生影响。前已述及，刑事智能预判与法官裁判在因果推理、事实认定和实践理性方面存在根本差异，这些差异不仅仅是技术层面的局限，更是造成认知冲突的潜在原因。换言之，刑事智能预判与法官裁判的认知冲突是二者认知模式差异的必然外化。在认知能力层面，法官可能受到智能预判模式识别的干扰产生认知偏差；在认知态度层面，法官则可能陷入笛卡尔式焦虑，在是否应该信任智能预判的问题上左右摇摆。双重层面上的认知冲突共同构成了法官的认知困境。

#### （一）加剧模式识别偏差

模式识别偏差是假设的来源，隐藏在所有推理背后，其会利用人们曾经经历过的模式来塑造当下对复杂现实的理解。人们以为是自己辨识出来的模式，实际上可能只是经自身确认的一种假说、一个正在讲述的故事的线索、关键角色的个性等。<sup>③</sup>这种偏差的效应就是使现实看上去比实际情况更简单、更连贯、更容易处理，其不必然导致案件的误判，却会使程序正义实质上面临巨大减损。刑事智能预判系统可能借由“光环效应”无意中强化了法官的固有偏见；通过虚假推演诱使法官陷入“故事陷阱”，对案情信息产生错误认识；令法官形成“后见之明”偏差，对预判结论产生莫须有的信心，进而在审查过程中掉以轻心。

##### 1. 光环效应固化裁判偏见

光环效应（Halo Effect），又称晕轮效应，指的是人们往往会因个体或事物的某一突出特征，对其整体作出偏颇的评价，也可以解释为使用基于不相关标准的评价来对某事或某人作出判断的行为。<sup>④</sup>比如，如果一个人外表得体、举止优雅，人们可能会下意识地认为他更聪明、更诚实。在司法审判中，光环效应无疑也会影响法官的判断，当判断具有负面含义时，例如“受教育水平较低的人更容易犯罪”，这种倾向则被称为“尖角效应”。

在刑事智能预判的应用中，这种认知偏见不仅可能被继承，还可能被放大和固化。当前实践中，多地法院开发的要素式智能审判很大程度上就面临着这种风险。此类系统通常将某类刑事案件分解为多个固定要素，并通过算法对这些要素进行量化分析和评估。<sup>⑤</sup>这种设计依赖于对要素的“标签化”处理，系统必须将复杂的案件事实简化为可定义的元素标签，例如初犯、累犯、自首、拒不认罪等。这些元素标签本应得到

<sup>①</sup>李桂林：《法律推理的实践理性原则》，《法学评论》2005年第4期。

<sup>②</sup>张欣：《生成式人工智能的算法治理挑战与治理型监管》，《现代法学》2023年第3期；胡铭、张传玺：《人工智能裁判与审判中心主义的冲突及其消解》，《东南学术》2020年第1期；马靖云：《智慧司法的难题及其破解》，《华东政法大学学报》2019年第4期。

<sup>③</sup>[法]奥利维耶·西博尼：《偏差》，贾拥民译，北京：中国财政经济出版社，2022年，第129页。

<sup>④</sup>See Guthrie C., Rachlinski J. J. and Wistrich A. J., Inside the Judicial Mind, 86 *Cornell Law Review*, 779 (2001).

<sup>⑤</sup>宁波市某法院的法官访谈，来自作者访谈记录。

法官审慎的综合考量，但由于智能预判缺乏实践理性的特点，其在基于历史数据学习时，某些负面标签可能会与高犯罪风险或低信用度等结论产生统计学上的强关联。当被告人被贴上一个或多个强烈的负面标签（如严重暴力犯罪、累犯）时，智能预判可能会作出整体的负面评价，预测较高的量刑区间或再犯风险。这意味着，即便其他要素（如正当防卫、有悔罪表现）是正面或中性的，也可能被突出的负面标签所掩盖，这正是光环效应的体现。

在人工将非结构化数据结构化输入算法数据库的过程中，被告人的年龄、性别、职业、外貌、受教育背景等标签特征均会被标记处理。<sup>①</sup> 智能预判系统高度依赖于历史数据，而这些数据往往承载着社会固有的偏见，例如倾向于认为高学历、白领职业的人再犯概率较低，而低学历、失业人员的再犯风险较高。越是智能化程度高的预判系统，越有可能受到光环效应影响，通过模式识别在标签之间自主发现这类“规律”，并沿用到对被告人的量刑预测和人身危险性评估中。无论这些偏见是正面还是负面，均存在以单一特征主导整体印象的倾向。在接收到智能预判结论后，本就存有相同偏见的法官将更容易巩固偏见形成先入为主的印象，而本身没有此类偏见的法官亦有可能疏忽大意接受了智能预判的错误导向，最终因被告人与案件无关的特征对其作出倾向性评价，干扰了判决的公正性。

## 2. 叙事陷阱误导事实认定

叙事陷阱（Narrative Fallacy）是指人类倾向于用因果关系构造连贯的故事来解释复杂的事件，即便这些事件本质上是随机的、无关的或受到众多未知因素影响。<sup>②</sup> 这种认知偏差使人们更愿意接受那些看起来合理的故事，而非充满不确定性的现实。正如纳西姆·塔勒布所言，“我们的头脑是一台非常了不起的解释机器，能够从几乎所有事物中分析出道理，能够对各种各样的现象罗列出各种解释。”陷入叙事陷阱的思维方式表现为在听到一个故事时，人们的自然反应是去极力寻找能证实它的各种元素，并且最终常能够找到这样的元素，根本不会想到一切可能只是偶然杂凑的结果。

缺乏因果推理能力和事实认定能力的刑事智能预判系统更易制造“合理但虚假的故事”，由于其推理过程并不是基于真正的因果关系，且无法实质性地深入分析审查证据，故而往往忽视了案件的复杂性与个体的特殊性，遗漏了事实经过的其他可能性，使法官对案件事实产生认知扭曲。<sup>③</sup> 智能预判倾向于将案件的复杂关系简化为一个线性叙事，其可能基于一系列孤立的事实构建一个看似合理的犯罪故事。举例来说，某一刑事案件的在案证据可证明如下事实：证人证言表明被告人与被害人之间存在纠纷，交易记录证明被告人在案发前购买了刀具，监控视频显示被告人向被害人捅刺数刀。对此，行为应当如何定性、是否存在防卫情节、是否有矛盾证据被忽略等问题，都需要根据个案具体情况展开全面查证。此类问题亦广泛存在于其他类型的刑事案件中，但智能预判的线性叙事倾向，却使其往往依据已知的碎片化事实拼凑成完整犯罪故事，直接生成相关结论。

这些结论虽然符合“故事逻辑”，却不一定符合“法律逻辑”。刑事审判需要遵循证据规则，排除合理怀疑，如果缺乏直接证据或间接证据不能形成完整严密的证据链条，则不能证明被告人有罪。刑事案件中的在案证据有时并不能完整还原案发经过，但刑事智能预判的模型设计却令其倾向于即时生成确定的结论，这一倾向便可能导致其机械化地推演出在案证据无法证明的虚构事实，进而提出错误的定罪量刑建议。法官在审查环节或许自认为已经进行了严格的事实核查（对事实的核查固然是必要的，因为不能排除智能预判会有编造事实的行为），<sup>④</sup> 但是，智能预判系统还可能会从正确的事实中推出错误的结论。当法官被看似逻辑连贯的完整犯罪故事影响，则可能忽略故事缺乏充分证据证明、无法排除合理怀疑等漏洞，导致作出错误裁判。

## 3. 后见之明遮蔽系统漏洞

---

①李本：《美国司法实践中的人工智能：问题与挑战》，《中国法律评论》2018年第2期。

②[法] 奥利维耶·西博尼：《偏差》，贾拥民译，北京：中国财政经济出版社，2022年，第14-16页。

③韩振文：《智能预判应用风险的技术性正当程序控制》，《中国社会科学报》2024年7月4日，第A04版。

④See Murray M. D., Artificial Intelligence and the Practice of Law Part 1: Lawyers Must be Professional and Responsible Supervisors of AI, 87 KBA Bench and Bar Magazine, 36(2023).

后见之明偏差(Hindsight Bias)是指当人们得知某一事件结果后,夸大原先对这一事件的猜测的倾向。<sup>①</sup>人们对过去的记忆是一个动态重构的过程,往往把后来得到的信息整合到过去的回忆中。<sup>②</sup>这就导致在知道一个实际上不可预见事件的结果后,人们会高估自己预测过去的的能力,倾向于认为这一结果“早该被预见到”或“是显而易见的”。这种认知偏差可能导致记忆失真,在回忆与重建内容时产生错误的结论。

刑事智能预判与法官裁判的认知差异,同样为后见之明偏差提供了滋生的土壤。随着智能预判介入刑事司法,法官一般是在参考智能预判系统给出的结果后,再作出最终裁判。这使得法官在接收到智能系统对案件事实的描述后,可能会下意识地调整自己原先的判断,以符合智能预判结论。这种偏差可能导致法官高估了还原案件发展轨迹的可预测性,且其作用方式常常是潜移默化的,法官很难凭借自身司法经验彻底避免。这也将导致法官后续的审查过程都是在智能预判结论基础上对案件事实的一种倒推演绎,疏忽了校正偏差的义务。

同样,当智能预判系统提供关于被告人的量刑建议和再犯风险评估等预判时,法官也会产生后见之明偏差,认为该评估或量刑“显然合理”,而意识不到最初可能仍存在怀疑的想法已经受到了智能预判结论这一“后来信息”的影响。倘若结论与法官凭自身直觉得出的结论大致相符,即使这只是一个巧合,也会使法官对该结论更加充满信心。这很可能减少法官对案件特殊性的考量,忽略了庭审中呈现的可能导向无罪或从轻处罚的情节。

## (二) 陷入笛卡尔式焦虑

多数司法境况中,面对智能预判系统生成的裁判结论,法官有可能存在“技术迷信”和“盲目排斥”两种不良倾向,<sup>③</sup>进而产生类似“笛卡尔式的焦虑”,即陷入相信真理与普遍怀疑两分式思维的矛盾之中:一方面基于自动化偏见,对聚焦的智能预判作出的参考结论持有乐观态度;另一方面又因算法厌恶倾向,对智能预判系统贸然给出的结论存有抵触怀疑,认为其与个案公正之间的本质关联其实并不完美。尽管自动化偏见与算法厌恶在理论上似乎相互矛盾,但在实际的人机交互场景中,两者往往交替或并存于同一位法官的思维中,构成“信任—怀疑—反复—陷入焦虑”的动态心理过程,这可能导致法官出现认知失调,陷入决策焦虑。

### 1. 自动化偏见

人们具有一种过于相信机器自动所作判断的认知倾向,尤其是在不确定状态下,容易对机器投以非理性的过度信任。<sup>④</sup>这种对自动生成结果的过度信任可能源于人们对于自身主观判断出错可能性的抵触:在司法系统中,人类法官的裁判可能受到个人能力、偏见、情绪等因素影响,作出不一致或错误的裁判。相反,智能预判作为客观工具,能够在算法驱动下稳定地输出预判结论,减少人类因主观不确定性产生的错误。法官可能因此过度信任自动化司法结论的可靠性和科学性,从而产生自动化偏见。

实践中把使用智能预判率作为质效考核的重要指标,评价法官对智慧办案系统的运用水平,<sup>⑤</sup>这也会加剧法官的自动化偏见,或引致其主体性出现自我庸化消解态势。在社会的数字化进程中,大量证据表明人们容易过度依赖智能决策系统等机器的辅助,包括法官在内的公众产生了一种“输出迷恋”,即计算机程序输出的结果不能被质疑,并被赋予了最终的技术权威。<sup>⑥</sup>这种对智能预判“客体式”的强烈依赖,使得法官决策变得脆弱,伴随着法律标准向技术标准的让步,便可能取代法官作为自觉自主的信息处理主体进行

---

<sup>①</sup>See Fischhoff B., Hindsight  $\neq$  Foresight: The Effect of Outcome Knowledge on Judgment under Uncertainty, *Journal of Experimental Psychology: Human Perception and Performance*, 304(1975).

<sup>②</sup>See Hawkins S. A., Hastie R., Hindsight: Biased Judgments of Past Events After the Outcomes Are Known, 107 *Psychological Bulletin*, 311(1990).

<sup>③</sup>谢登科:《人工智能证据的类型分析与规则建构》,《学习与探索》2025年第3期。

<sup>④</sup>See Robinette P., Howard A., Wagner A. R., *Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed?*, Springer, 129(2017).

<sup>⑤</sup>郑智航:《“技术—组织”互动论视角下的中国智慧司法》,《中国法学》2023年第3期。

<sup>⑥</sup>高童非:《数字时代司法责任伦理之守正》,《法制与社会发展》2022年第1期。

的裁量心证判断。自动化偏见亦与“机器权威”现象相似,容易引发严重的“偏见回声”效应,<sup>①</sup>即法官预判中的偏见<sup>②</sup>与智能预判结论之间发生认知共振,从而固化了法官的认知偏误。

由此,便可能引致一种“人机合谋”的极端形态,<sup>③</sup>即法官倾向于凭法感盲目偏信智能预判系统输出的结论是科学正确的,甚至在庭审过程中仍为智能预判结论所左右,通过调整自身认知以匹配预判结论,而不对其进行交互理性的验证反思,最终的裁判结果与智能预判结论趋同甚至等同。<sup>④</sup>尽管理性审慎的法官对预判结论往往会保持“将信将疑”的态度,但“人机合谋”的风险却并非耸人听闻。当刑事智能预判系统不断优化升级,无疑会筛选出采纳率更高的算法模型,其表面看来越是显得操作精细准确,法官越有可能固化放大其原有的自动化偏见,产生一种神明裁判式的非理性崇拜。这种倾向将有违司法裁判的亲历性原则,以至于出现算法独裁而被算法锁定,实际裁判最终沦为技术附庸。

## 2. 算法厌恶

自动化偏见使法官倾向于高估智能预判的准确性与权威性,然而,一旦智能预判系统对于某个案件的预判结论出现偏差或错误,法官的信任基础也可能迅速动摇,出现“算法厌恶”(Algorithm Aversion)。算法厌恶表现为一种对算法的偏见性评估,即人们对算法持消极、不信任态度,倾向于拒绝来自算法的建议或推荐。<sup>⑤</sup>这种机制作用于法官认知模式中,则表现为对刑事智能预判系统整体能力产生质疑,降低对预判结论的参考意愿,甚至抵制其使用。法官对自身认知的控制感以及对算法不透明性的抗拒都会触发算法厌恶机制。

根据元认知监控理论,个体在进行认知活动的全过程中,会将自己正在进行的意识活动作为意识,不断对其进行积极、自觉的监视、控制和调节。<sup>⑥</sup>这种监控反映到司法活动中,就表现为法官的裁判结论需要以明确且可靠的认识为基础,必须能够清晰认识到自己的裁判结论是如何形成的。面对刑事智能预判系统生成的结论,法官同样会开启元认知监控,通过反思自己的思维方式、推理过程和判断标准,对智能预判的可信度进行评估。法官经过多年专业训练,积累了复杂的法律思维,他们往往清楚自己在案例分析、证据评估和法律推理中的方式与人工智能截然不同。尤其是年纪较大的法官长期依赖传统的办案模式,不适应新的技术工具,更加难以信任智能预判系统的处理结果。<sup>⑦</sup>智能预判形成过程排斥了其他主体的参与,削弱了法官的判断能动性,这很大程度上导致法官在评估其所作结论时存在抵触情绪。

算法的不透明性亦会导致法官对智能预判结论产生类似的倾向。由于当前尚不具备相应的算法解释技术,刑事智能预判系统中深度学习模型的运行具有不透明性,通常难以通过自然语言被解释。<sup>⑧</sup>具体而言,强大的智能预判系统离不开机器学习技术,机器学习的本质就是从数据中找规律,再利用规律作预测。机器学习是计算机主动学习数据中的规律,而不是由研究者人为地提前输入规律。<sup>⑨</sup>需要注意的是,机器学习中所寻找的规律,有的容易被人类所理解,例如决策树模型和回归模型可以清晰地展示自变量对于结果变量影响的强度和方向;有些规律却难以理解,甚至无法言说,比如人工神经网络模型(深度学习),其在预测效果上有着非凡的表现,但整个预测过程近似“黑箱”,很难解释每个自变量对结果变量的影响,即其决策过程往往缺乏可解释性。当法官在使用刑事智能预判系统时无法理解或解释系统的预判结论,就可能导致他们对该结论缺乏信任,甚至产生厌恶抵触。

由于自动化偏见和算法厌恶并不是稳定的性格特质,而是情境驱动的心理状态,因此当二者同时作

①李学尧:《大语言模型应用中的司法偏误与认知干预》,《政治与法律》2025年第5期。

②韩振文:《智能时代法官预判确定性的功用、契机及其困境》,《上海政法学院学报(法治论丛)》2023年第2期。

③衣俊霖:《论人机协同审判的信任构建》,《中国法学》2025年第2期。

④周芷妍:《数字时代法官预判确定性与认知偏差控制》,《智慧法治》集刊2024年第1卷。

⑤See Dietvorst B. J., Simmons J. P., Massey C., Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err, 144 *Journal of Experimental Psychology: General*, 114(2015).

⑥周爱保主编:《认知心理学》,北京:清华大学出版社,2021年,第254页。

⑦玉环市某法院的法官访谈,来自作者访谈记录。

⑧周翔:《算法可解释性:一个技术概念的规范研究价值》,《比较法研究》2023年第3期。

⑨刘友华:《算法偏见及其规制路径研究》,《法学杂志》2019年第6期。



用,便会导致法官产生一种双重心理悖论。当智能预判系统既被感知为“高效工具”又“可能犯错”,认知冲突便引发法官思维上的反复,感到无所适从,进而陷入决策中的“笛卡尔式焦虑”状态。

#### 四、刑事智能预判与法官裁判的认知协调

以上对刑事智能预判形成机理与人类认知模式的差距进行了初步勾画,并指出了其影响法官对案件认知判断的潜在风险,而当前对刑事智能预判的质疑,不仅源自其本身的系统性风险,更根本的挑战在于其对程序正义,特别是“主观程序正义”的冲击,技术引发的风险呼唤着传统正当程序原理的技术化重塑转型。<sup>①</sup>而在主观程序正义视角下,司法是否正当不仅取决于程序设计本身是否符合法定规则,更取决于司法参与者是否感知、体验到程序的公正。<sup>②</sup>故在推动技术性正当程序规则构建、基于案件复杂度动态调整智能预判介入深度的同时,也必须引入参与感、可接受性、可质疑性等评价维度,实现形式正义与感知正义的统一。与此同时,关于消除偏差的研究发现,让决策者依靠自己的力量去消除偏差是非常困难的。因此,与其期望法官在高度复杂的认知负荷下自发诊断并消除智能预判偏误,不如通过构建以主观程序正义为导向的技术性正当程序,优化其决策流程与制度环境,实现对法官的“程序赋能”,从而在认知上实现技术协作而非技术依附。

##### (一) 限定智能预判应用边界

在强人工智能还未成为现实的当下,智能预判系统在因果推理、事实认定以及价值判断方面的缺陷将在较长一段时期内延续,可能产生的错判误判风险也难以在短期内消除。有鉴于此,根据“场景化正当程序”理论,<sup>③</sup>有必要注意智能预判应用的具体业务场景,通过类型化实现制度研究精细化,根据案件类型和复杂度,构建“应该用、可以用、选择用、禁止用”的动态分类分级机制来实现相对可控可信的刑事智能预判应用。<sup>④</sup>具体而言,应当严格限定刑事智能预判的应用范围和边界,依据简单刑事案件与疑难刑事案件的复杂度差异,调整智能辅助决策应用的环节。在规则制定上,明确刑事智能预判系统的准入门槛,设定不予适用的案件类型“负面清单”,以期有效避免智能预判对法官独立决策质量的减损。

在不同案件类型中,智能预判对法官认知的影响广度和强度并不一致,法益保护需求的程度也有所差异。基于技术上的可行性,具有深度学习能力且证据推理完全自动化的智能预判系统,其应用场景通常限定于处理格式化、标准化的数据信息。放到刑事领域中,首先,对于案件数量巨大同时事实要素清晰、行为模式单一、证据种类较少、法律关系简单的案件类型,比如醉酒型危险驾驶案件,要在保证公正裁判的同时兼顾效率,就应当借助智能预判系统逐步探索问题的解决途径。<sup>⑤</sup>其次,对于适用简易程序、速裁程序等审理的以及缓刑、减刑、假释等程序简便的案件,也可以适用智能预判进行辅助审理。不同于复杂疑难案件中蕴含较多的信息要素,此类案件中的事实情况和证据类型往往较为简单明了且程序简单,<sup>⑥</sup>智能预判在介入时不存在太多的技术难题,有助于保障算法决策的精确性。对于此类案件,算法能够较为充分完整地处理数据,发挥智能预判高效审查的作用,实现简审快审。

与之相对,基于智能预判在事实审查、价值判断等方面的技术局限,对于使用普通程序合议庭审理的以及被告人拒绝认罪或辩护人作无罪辩护的案件,则一般不宜适用智能预判,<sup>⑦</sup>以防加剧法官认知偏见。当然,考虑到智能预判的证据梳理、类案推送、裁判偏离度预警等功能仍可能起到辅助作用,因此不必然排

① 郝龙飞:《韧性 with 规划:刑事司法算法决策的演进逻辑》,《北方法学》2024 年第 1 期。

② 苏新建:《程序正义对司法信任的影响——基于主观程序正义的实证研究》,《环球法律评论》2014 年第 5 期。

③ 冯建鹏:《基于主观程序正义的技术性正当程序——我国算法治理的一种路径》,《法制与社会发展》2025 年第 1 期。

④ 王毓莹:《算法裁判风险的形成归因与化解》,《法律适用》2024 年第 5 期。

⑤ 王伟、石改军、赵天培:《醉酒型危险驾驶罪智能化刑罚裁量体系构建——基于对 2001 份相关判决的考察》,《法治文化》集刊 2025 年第 1 卷。

⑥ 丰怡凯:《人工智能辅助量刑场景下的程序正义反思与重塑》,《现代法学》2023 年第 6 期。

⑦ 蔡立东、郝乐:《司法大数据辅助审判应用限度研究》,《浙江社会科学》2022 年第 6 期。



除刑事智能预判在此类案件中的适用,但在选择性适用时必须坚持对智能预判结论的有限参考原则,<sup>①</sup>在适用时保持必要的谨慎。在具体的流程设计上,还要对智能预判系统具体功能的适用环节作出明确规定,包括严格限制证据分析、量刑建议、人身危险性评估等信息进入法官认知视野的时间节点。例如为避免量刑预测影响法官心证,可将此类推送环节延迟,在法官已独立形成量刑意见后再进行参考匹配,尽可能减少其对法官认知的负面干扰。

最后,需要进行价值判断的疑难复杂案件,则是算法决策适用领域的核心禁区与负面清单。这是考虑到智能预判系统在法律解释、证据能力的实质性评价、自由心证及经验法则的运用、内心确信的形成等主观性较强的事项上,存在难以回避的认知障碍及作用局限。<sup>②</sup>而疑难复杂案件裁判的证明须从情理角度对证据能力和证明力进行实质性审查与综观式验证,蕴含着诠释评价、自由裁量等辩证思维活动,目前的智能预判系统难以对这种裁判权运作进行算法表达。可以说对于刑事疑难案件,富有经验的法官都很难判定处理,更不用期待人工智能能够实现精准裁量模型的构建,生成合理准确的算法自动化决策。当然刑事疑难案件会随条件变化可能转化为简单案件,所以负面清单应是动态设定的。特别需要指出的是,刑事智能预判应禁止适用于定罪环节。这是考虑到定罪涉及对案件事实的认定和法律性质的判断,具有难以动摇的政治哲学支撑,<sup>③</sup>高度依赖人类主观能动性和复杂的价值评判,智能预判尚无法有效胜任。

## (二) 优化预判结论可解释性

智能预判结论的可解释性意味着其生成过程和输出结果应当能够以某种方式向人类合理解释。在主观程序正义框架下,解释不仅是一种技术说明,更是对司法参与者正义感知的回应与司法信赖的前提。<sup>④</sup>智能预判应提供公开的理由阐明特定算法如何生成预判结论,以打开算法黑箱证明裁判的正当性。<sup>⑤</sup>尽管典型的算法黑箱模型,如神经网络模型的可解释性较弱,但需要指出的是,这类算法并非完全不可解释。<sup>⑥</sup>目前以 ChatGPT、Deepseek 为代表的大语言模型,可以通过“深度思考”选项更具体地呈现其结论生成的逻辑,意味着生成式人工智能的部分预测内容是具备现实可解释性的。因此,应当着力将“黑箱”模型转化为“灰箱”,以优化智能预判的可解释性。

关于算法解释技术的分类方式众多,包括全局解释和局部解释、透明化和事后归因等,但究其本质其实大同小异,从字面含义上看,将两种解释类型称作“内部解释”和“证明解释”较为容易理解。<sup>⑦</sup>“内部解释”即解释算法的本身,旨在使算法的整体运作过程能够为人们所理解。其基本的思想是尽可能阐明模型的输入特征和输出结论间的相关性,<sup>⑧</sup>用于说明该复杂黑箱模型的结构。“证明解释”则是解释算法输出的结果,其无意于揭示算法内部的运行逻辑,而是另行设计一个模型同步生成对输出结果具有针对性的个别化、具体化解释,<sup>⑨</sup>解释的方式可以是一段自然语言描述或是相近似的案例等。

以上两种解释方式的角度和技术各有侧重,内部解释能更完整详细地呈现算法模型的运作过程,证明解释则不涉及算法构造本身;但在用户体验上,证明解释相对通俗易懂,对不精通技术的广大人群而言更为友好。考虑到刑事智能预判系统的说理辅助价值主要集中于为已经得出的裁判结论寻找充分的证成理

① 戎静:《“预测正义”能否预测正义?——基于法国司法大数据预测应用的考察与启示》,《中外法学》2023年第5期。

② 高翔:《智能司法的辅助决策模型》,《华东政法大学学报》2021年第1期。

③ 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》,《政法论坛》2021年第5期。

④ 郭春镇:《感知的程序正义——主观程序正义及其建构》,《法制与社会发展》2017年第2期。

⑤ See Leah Wissner, Pandora's Algorithmic Black Box: The Challenges of Using Algorithmic Risk Assessments in Sentencing, 56 *American Criminal Law Review*, 1831(2019).

⑥ 刘东亮:《技术性正当程序:人工智能时代程序法和算法的双重变奏》,《比较法研究》2020年第5期。

⑦ See Lipton Z. C., The Mythos of Model Interpretability: In machine Learning, the Concept of Interpretability Is Both Important and Slippery, 16 *Queue*, 31(2018).

⑧ See Rudin C., Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 1 *Nature Machine Intelligence*, 206(2019).

⑨ See Lundberg S. M., Lee S. I., A Unified Approach to Interpreting Model Predictions in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768 (Ulrike von Luxburg ed., Curran Associates Inc. 2017).

由,助力提高文书释法说理的质量,<sup>①</sup>对算法的解释不需要达到让法官理解算法应用技术细节与运作逻辑的程度。由此,可以尝试对智能预判采用一种折中式的解释方法,即对算法可解释性采取适度优化。<sup>②</sup>通过对模型进行大幅度的精炼和概括,选择可视化、交互界面、示例解释等“用户友好”的算法解释进路,例如用降维技术达成法官可理解的简单可视化结果。

在具体开示内容上,可以呈现有关算法技术架构方面的适度数据信息,比如所用数据标注净化的要素、数据集选择范围与主要特征、训练模型的算法类型等。在量刑建议和人身危险性评估中,算法的功能和推理过程应得到充分解释。例如,系统如何结合案件历史数据,通过特定的变量(如犯罪类型、累犯记录、社会背景等)来确定量刑范围和人身危险性。此类功能解释应侧重于算法的证明解释,即展示每种指标的功能权重、机器定义的特定案例决策规则、对量刑区间划分和危险性评估起参考辅助作用的信息等,将这些内容以文字说理的方式呈现给法官。只有通过增强智能预判结论的可解释性,保证结论有推理分析的相对适度透明,才能实现智能预判与法官的交互理解,减缓法官对智能预判结论瞬时产生的怀疑不安,以保证智能预判结论具有可验证性,并在法官认可的正当范围内合规运行。

### (三) 构建法官分类参考机制

在刑事智能预判介入裁判活动的背景下,如何保持法官的裁判主体性并对智能预判所提供结论进行可感知的审查与解释,使当事人与社会公众感知到法官实质性地参与到对争议的解释和说明理由中,正是实现主观程序正义的关键。<sup>③</sup>为保障法官在裁判过程中的认知思维不受智能预判负面干扰,应当依据上文“应该用、可以用、选择用、禁止用”的案件类型动态划分,实行法官在不同案件中对智能预判结论的分类参考机制。发挥法官作为裁判主体的能动作用,赋予法官在参考匹配智能预判问题上的选择权利,同时要求其承担采纳适用时的审查和论证义务。这样既可明晰人机责任分配、增强法官审判伦理责任,避免决策权在事实上让渡的风险,也可最大限度防止法官的认知判断产生扭曲。

《最高人民法院关于规范和加强人工智能司法应用的意见》明确指出,要“坚持对审判工作的辅助性定位和用户的自主决策权”。技术规训措施及其结果应主要作为规范司法行为的一种参考,诸如证据分析、量刑建议以及人身危险性评估不应成为支配司法权力的直接根据。<sup>④</sup>因此,需要明确法官享有免受智能预判约束的权利。首先,除“应当适用”刑事智能预判的标准化简单刑事案件外,法官应当有权主动放弃智能预判结论与待决案件相比照的匹配环节。鉴于系统很难对默会知识进行建模推理,面对层出不穷的疑难案件,为避免法官的裁判权遭到算法决策的介入和稀释,法官可根据案件情况和实际需求决定是否参考智能预判结论,或可选择快捷关闭个性化的预判推送。其次,在参考了智能预判结论的情况下,法官亦有权选择不予采纳算法完全自动化处理作出的决定,这也是技术性正当程序的重要体现,即突出人的主体性价值与存在意义。同时,智能预判既然只是一种参考性意见,那么不予采纳的理由便属于法官自由心证的部分,无需加重法官说理负担。

相反,如果法官选择采纳智能预判所作的结论,就同时负有对智能预判结论的审查和论证义务,对裁判行为和结果质量承担责任。<sup>⑤</sup>在审查环节,法官应结构化地监督、复核预判结论是否为算法有效执行的结果。对于智能预判生成的量刑建议或者法律文书,法官必须审阅其中影响定罪量刑的关键因素,对其运行的关键环节和关键技术达到熟练理解的程度,对相关参数的自由设定以实现对智能预判结论的主动把关。<sup>⑥</sup>在此基础上自主对生成结论进行校验、修正优化,保证生成过程的数字在场,以此强化法官对智能预判的监督审查责任,实现安全可信技术伦理与法官认知推理能力相对齐。

在论证环节,法官有必要进行适度释法说理,包括智能预判认定事实、适用法律的根据理由。对直接影响诉讼过程及结果的程序性事项给予必要反馈回应,才能彰显裁决的合理可接受性。法官仅仅利用和整

①周翔:《司法人工智能对裁判说理的辅助价值和实现路径》,《法学杂志》2024年第1期。

②苏宁:《优化算法可解释性及透明度义务之诠释与展开》,《法律科学(西北政法大学学报)》2022年第1期。

③罗英:《数字技术风险程序规制的法理重述》,《法学评论》2022年第5期。

④左卫民、潘鑫:《通过技术规训司法:进步与挑战》,《法学评论》2023年第4期。

⑤雷婉璐:《智能辅助审判技术下法官问责难题的破解路径》,《哈尔滨工业大学学报(社会科学版)》2022年第2期。

⑥龚善要:《人工智能司法应用的实践审思与完善》,《国家检察官学院学报》2023年第5期。

合法律语义知识来认知智能预判结论是不够的,还有必要对输出的预判结论提供有意义的解释证成,以确保其结论在法官主导下进行最终验证。尤其是对于只能“选择性适用”刑事智能预判的案件,通过书面理由说明是如何考虑和应用智能预判所提供的参考意见,以备上级审查和利害关系人寻求救济。<sup>①</sup>由此实现人机分布式道德责任分配,倒逼法官审慎对待参照智能预判结论,最终建立起裁判受众对人机交互沟通决策的普遍接受与信任生态。

## 结 语

在坚持繁简分流机制改革、优化司法资源配置的背景下,智能预判的逐步深入应用无疑顺应了人民群众对高效便捷解纷的司法需求,刑事纠纷的解决同样有必要采取数字化措施来协同提速增效。但在人工智能奇点尚未来临的今天,刑事智能预判系统并不真正具有人类智能,其中凝结的更多是算法设计者和开发者辛勤的汗水,指望智能系统一劳永逸地解放法律人,显然还为时过早。面对当前弱人工智能带来的风险与挑战,法官有必要更新传统的裁判认知思维,在实践中逐步增强理解算法和解释算法的能力,充分发挥人工智能在辅助决策方面的功用。为此需要通过技术性正当程序的规制以及算法解释技术的完善,保证法官既不会迷信盲从自动化决策,又不至于对预判结论完全失去信任,而是能够客观地进行参考审查。随着智慧司法进程不断加快,算法自动化决策质量及配套规则设计都将持续提高和完善,人工智能深度融入司法决策也将愈发能被裁判受众所理解与信任。

## Cognitive Conflicts and Coordination Between Criminal Intelligence Prediction and Judges' Judgment

HAN Zhen-wen<sup>1</sup>, ZHOU Zhi-yan<sup>2</sup>

(1. School of Law, Zhejiang Gongshang University, Hangzhou 310018, China;

2. School of Criminal Law, East China University of Political Science and Law, Shanghai 201620, China)

**Abstract:** With the rapid iteration and advancement of generative artificial intelligence, the deep integration of artificial intelligence into the criminal justice system has become an increasingly prominent trend. However, it is important to recognize that the operational logic of current criminal intelligence prediction still diverges significantly from cognitive process of judges' judgment. These systems exhibit key limitations: an inability to reason about causality, difficulty in establishing facts of the case, and a lack of practical rationality. Such technical constraints may lead to cognitive conflicts between the predictive outputs of intelligent systems and the judicial reasoning of human judges. On the one hand, pattern recognition biases may distort judges' decision-making; on the other hand, it makes judges fall into a bipolar situation of blind obedience or suspicion of the intelligent prediction result because of the ubiquitous cognitive tendency. Therefore, it is necessary to adhere to the regulatory philosophy of technological due process by, based on the complexity of criminal cases, dynamically delineating and adjusting the application boundaries of intelligent prediction, optimizing the interpretability of the intelligent prediction outcomes, and establishing a classified and graded reference mechanism for judges', so as to reconcile the cognitive conflict between criminal intelligence prediction and judges' judgment as much as possible.

**Keywords:** criminal intelligence prediction; judges' judgment; cognitive conflict; technological due process

[责任编辑 治平]

---

<sup>①</sup>刘金松:《数字时代刑事正当程序的重构:一种技术性程序正义理论》,《华中科技大学学报(社会科学版)》2023年第2期。