

论人机协同审判的信任构建

衣俊霖*

内容提要 随着人工智能在我国智慧司法建设中的普及推广,司法审判逐步呈现出“法官为主、机器为辅”的人机协同审判格局。目前,司法人工智能暂时消隐于法官背后,审判信任被置换为对法官人格、法院制度或人工智能算法的信任。人格信任奠基于法官作为“完人”的应然预设,与实然的法官“常人”之间存在现实差距。法院组织以司法责任制作约约束人机互动的主要制度安排,但在预测审判逻辑的影响下,存在盲目采纳机器建议的“人机合谋”风险。算法信任机制尽管能发挥局部的信任补救功能,但囿于“主体—工具”的二分思维,难以对法官与算法的协同过程进行约束。应当以信任者与信任对象的沟通互动为出发点,对人格信任、制度信任和算法信任进行反思性整合,将司法人工智能从内部办案工具拓展为外部沟通工具,推动人机协同法律议论的充分展开,为人机协同审判奠定更为坚实的信任基础。

关键词 人工智能 人机协同审判 审判信任 法律议论

目次

- 一、人机协同审判的信任主题
- 二、审判者人格信任的现实局限
- 三、审判机关制度信任的异化风险
- 四、人工智能算法信任的局部补强
- 五、基于沟通议论的整合式信任构建
- 六、结语

* 上海交通大学凯原法学院助理教授,法学博士。本文系2023年度国家社科基金青年项目“中国特色公共数据市场化运营的法理基础与制度构建研究”(项目批准号:23CFX018)的阶段性成果。

一、人机协同审判的信任主题

近年来,人工智能技术蓬勃发展,给人类社会带来一系列机遇和挑战。在海量语料、丰富算力的推动下,基于神经网络、深度学习的大语言模型,为人工智能技术带来了颠覆性革新,机器算法早已悄然跨越图灵测试的门槛。随着世界各国智慧司法建设的稳步推进,人工智能技术不断渗入、赋能法院司法审判工作,涉足审判执行、诉讼服务、司法管理等多个方面。我国自“十三五”时期起便在智慧司法领域进行布局,人类法官与人工智能深度融合的协同审判模式正逐渐成为司法审判工作的新常态。

(一) 人工智能介入司法审判的发展分期

人工智能对司法审判的介入,与计算机科学技术的突破性发展密不可分,且呈现出与颠覆性革新事件相关联的断裂现象。透过对技术断裂点的观察,大致可以划分出人机对抗和人机协同两条主线。前者聚焦于机器算法能否替代人类法官的问题,后者强调人类法官与机器算法在司法审判中的交互协作。当然,两者并不是非此即彼,而更像对位交织的两个声部,但仍可借助断裂点以辨识出特定时期学术争鸣的主旋律。

第一个断裂点发生于2016年,开启了对审判主体性人机对抗问题的热烈讨论。谷歌公司于当年推出的AlphaGo战胜了著名棋手李世石,成为科技界的轰动事件,同时触发了社会各界的广泛反思——究竟哪些领域是人工智能仍没有能力或不应涉足的人类阵地。在此大背景下,法学界关于人类法官与人工智能的“人机对抗”之辩自此拉开帷幕。^①一定程度上,法官成了代表人类智慧与人工智能进行新一轮对弈的棋手,只不过棋盘转变为司法审判这一颇为敏感而复杂的领域。此外,广受关注的美国“卢米斯诉威斯康星州案”也在2016年终审宣判。在该起案件中,人类法官的审判主体地位并没有变,但法官实质上是否已过度依赖机器算法的问题引起了不少担忧。因此,在人机对抗的主战场外,对人机协作潜在风险的反思也埋下了伏笔。

第二个断裂点发生于2022年,“法官为主、机器为辅”的协同审判模式逐渐成为通说共识。在科技领域,ChatGPT领衔的大语言模型技术在当年取得突破性进展。长期以来,人工智能处理人类自然语言文本的效果并不令人满意,妨碍了其在日常工作生活中的普及推广。而大语言模型彻底改变了这一现状,普通公众可以借助情景对话方式快速上手使用。由此,人工智能成为人们在软件研发、教学科研、内容设计等各领域的协作搭档,并开始出现“遇事不决,先问AI”的有趣景象。围绕审判主体性的人机之争也在当年尘埃落定,“法官为主、机器为辅”的人机协同审判模式在我国司法实践中得以确立。相关标志性事件是,《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

^① 参见潘庸鲁:《人工智能介入司法领域的价值与定位》,载《探索与争鸣》2017年第10期;郑曦:《人工智能技术在司法裁判中的运用及规制》,载《中外法学》2020年第3期;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》,载《政法论坛》2021年第5期。

〔2022〕33号)于2022年12月发布,集中回应了2016年以来人机对抗的相关争议。至此,人类法官的审判主体性基本得到巩固,更多研究开始转向法官与机器的协同问题。^②

需指出的是,我国对人机对抗、人机协同的讨论至少可追溯到21世纪初的“电脑量刑”。2004年,山东省淄博市淄川区法院引入电脑辅助量刑技术,可根据犯罪情节自动计算刑期。该尝试一时引起广泛热议,终因争议较大未能得到普及推广。^③ 受限于彼时的技术条件,所谓电脑量刑不过是基于量刑表格的简单查询和算术,远远不能与如今的ChatGPT、DeepSeek等大语言模型相提并论。不过,淄川法院当年也面临机器可能取代法官的质疑,而其所作的回应亦令人颇感熟悉——电脑只是辅助工具,法官才是审判主体。^④ 可见,关于人机对抗与人机协同的早期思考至今仍发挥着影响。

(二) 人机协同审判面临的信任挑战

从早年的电脑量刑,到如今基于大语言模型的“未来法官助手”^⑤,学术研究许多时候似乎是对技术里程碑事件的跟进与回应。技术革新使人们短时间内聚焦特定问题,但新的技术又随时抛出新问题,导致相关研究呈现些许跳跃甚至断裂。

然而,断裂的背后存在一条连续的逻辑主线,贯穿其中的主题是审判信任。信任一般是指对个体、组织或系统所实施行为的可靠性的信心状态。^⑥ 审判信任作为信任的下位概念,是指社会大众对于司法审判活动的信心和信赖。信任具有简化复杂性的作用:信任程度高时,人们会减少对信任对象的干预控制;而信任度低时,则可能采取控制行为以提升结果的可预期性,或回避与信任对象进行接触。就司法审判而言,审判信任的构建不仅有助于维护社会秩序和法律权威,还能促进裁判结果的有效执行。

在审判信任的观察视角下,有关人工智能介入司法审判的一系列探讨,从更深层次上都可被视为对信任问题的讨论。无论21世纪初有关电脑量刑的争议,还是2016年以来有关“AI法官”的审判主体性之辩,都体现了人们对机器算法介入审判——甚至窃取法官审判权——的不信任感。即便在“法官为主、机器为辅”的协同分工形成之后,人们对司法审判活动的信任显然也并非坚如磐石。^⑦ 从更广阔的视野来看,人机协同审判的信任风险与目前人工智能引起的信任危机密切相关。算法黑箱、算法偏见等一系列社会热点问题,在深层次上揭示出人工智能系统的信任困境。而人工智能对司法审判的介入,触动到人们内心深处对“人类被机器审判”的恐惧。可以说,人机协同审判的信任问题恰好处于这场信任危机的前线。

(三) 寻找人机协同审判的信任基础

审判信任并非人工智能时代的新问题,其在我国有时呈现为“司法公信力”的叙事

② 参见丁晓东:《人机交互决策下的智慧司法》,载《法律科学》2023年第4期;丰怡凯:《人工智能辅助量刑场景下的程序正义反思与重塑》,载《现代法学》2023年第6期。

③ 参见季卫东:《电脑量刑辩证观》,载《政法论坛》2007年第1期,第126-127页。

④ 参见萍梅:《电脑量刑是与非》,载《政府法制》2004年第24期,第5页。

⑤ 丁国锋:《苏州法院运用AI打造“未来法官”》,载《法治日报》2024年4月10日,第1版。

⑥ 参见郑也夫:《信任论》,中信出版集团2015年版,第1-16页。

⑦ 参见张凌寒:《智慧司法中技术依赖的隐忧及应对》,载《法制与社会发展》2022年第4期,第186-188页。

话语。司法公信力通常是指法院的信誉和口碑,即法院的行为、表现或成就所引起的公众评价。审判信任和司法公信力存在相互关联,但亦存在微妙而重要的差异。公信力是从作为被信任对象的法院角度而言的,而信任则强调从信任者的角度来观察信任对象。故此,审判信任是一个涵义更为丰富的概念,所涉及信任对象的范围也更加广阔,可以将法官个体、法院组织、司法人工智能尽皆纳入研究视野。

以审判信任为切入点,可以从“人工智能+司法审判”相关研究文献中提炼出三个信任类型以及相应的信任构建模式。其一,将审判信任问题予以化简,置换为对审判者之专业能力与道德品质的人格信任。其二,将审判信任寄托于对审判机关的制度信任,特别是对司法责任制、审判监督等制度的信任。其三,通过完善司法人工智能的准确性、可解释性、可问责性等指标,借助算法信任来夯实审判信任的基础。基于对前述三种类型的排列组合,又可以形成其他信任构建方式,使人工智能时代的审判信任呈现出异常丰富的内涵和多元化的构建进路。

通过对前述三类信任构建模式的批判与反思,笔者试图刻画审判信任得以建立和维系的核心逻辑,揭示审判信任在人工智能时代的变与不变,并在此基础上,提出一套基于法律沟通议论的整合式信任构建理论,为人机协同审判奠定更为坚实可靠的信任基石。

二、审判者人格信任的现实局限

社会大众对司法审判的信任在很大程度上与对法官的信任密不可分。作为法院与社会大众之间的一扇窗,法官因其专业能力、道德品质以及公正无私的品格而赢得外界的尊重和信任。随着人工智能应用于司法审判,首先浮现出来的问题是:人们是否像对人类法官那样信任人工智能。更进一步的追问是,在人机协同审判模式下,对人类法官的信任能否构成审判信任的坚实根基。

(一) 审判主体性的人机对抗

在人工智能发展的早期阶段,机器算法无法像人类法官一样进行复杂推理工作,所以对于“机器法官”的讨论更多处于理论想象层面。近年来,随着大语言模型的迅猛发展,人工智能已能够产生逻辑紧密、表述清晰的自然语言文本。因此,人工智能在法律领域的应用,成为人们普遍寄予厚望的一个落地场景。人工智能代表一种异于人类的智能运作逻辑,被置于人类智能与机器智能的二元对抗图式下加以审视。在这一背景下,人工智能是否会替代法官而成为审判主体,一度成为法学界的研究热点。

在审判主体性的争夺上,人类法官和人工智能以颇为抽象、简化的面目进行同台竞技。法官作为人类价值理性的代表,被浪漫化地视为专业知识和道德品格上的“完人”。在思维创造性^⑧、默会知识和实践理性^⑨、社会经验等方面,人类法官具有人工智能所无

^⑧ 参见潘庸鲁:《人工智能介入司法领域路径分析》,载《东方法学》2018年第3期,第114页。

^⑨ 参见前注^②,丁晓东文,第65页。

法比拟的优势。同时,完人法官具有同情心、同理心,可以在审判过程中实现情理法的融合。相比之下,人工智能则在效率、一致性、中立性等方面具有优势,可以快速处理大量案件信息,缩短审理时间、提升司法效率。算法亦不受个人情感和利益的影响,可以基于类案裁判历史数据推动裁判尺度的统一。不难发现,人类法官与人工智能的比较,被置于公正与效率、价值理性与工具理性的对立语境之下。

目前,人类法官取得了审判主体性之争的第一回合胜利。提倡价值理性、限制工具理性、坚守人类主体性,成为审判主体性之争的通说。具体而言,反对人工智能替代人类法官的理由主要包括:其一,人工智能难以妥善处理价值判断问题;其二,人工智能的性能主要由训练数据决定,故可能受到数据偏误的影响而产生偏见或错误;其三,人工智能对简单的类型化案件比较准确,但难以在复杂案件中实现个案正义。此外,有学者指出,人们之所以反对算法取代法官,并非由于审判准确性、公正性等方面的差异,而是源于主观的“感知正义”匮乏。^⑩总而言之,由于人们不信任人工智能可以胜任审判者的角色,人类法官的审判主体地位因而得以保全。

(二) 人机协同审判的法官人格信任背书

“法官为主、机器为辅”的人机协同审判模式,已成为理论和实务界的共识。具体来说,法官是审判活动的最终裁判者,对审理结果承担司法责任;与此相应,人工智能不得取代法官,其运行结果仅供参考。^⑪当然,尽管机器不能取代法官,但其作为审判辅助工具的应用前景是颇为光明的。在法院“案多人少”的压力下,人工智能在提升审判效率、加强司法管理等方面确有重要价值。^⑫我国法院已明确了司法人工智能作为辅助审判工具的功能定位。^⑬在坚持法官审判主体性的前提下,人工智能对法院审判的介入已颇为普遍。案件审理时,先由人工智能进行“粗加工”,再由人类法官实施更加精细的审理,这条兼顾正义与效率的人机融合进路似乎已呈现在眼前。

对法官审判主体性的坚持,在很大程度上将人机协同审判奠基于法官人格信任背书。所谓“人格信任”,即对“另一个人”的信任,他人此时作为有序的、人格化的行动者而成为被信任对象。^⑭对法官的人格信任,主要涉及专业能力和道德品质两个维度:在专业能力维度,信任法官能胜任甄别司法人工智能相关风险的工作,并纠正其中的偏误;在道德品格方面,信任法官决定是否采纳机器辅助建议时,可以公正无私地排除个人利益等无关因素。目前,理论和实务界把“信任票”投给了人类法官,不信任人工智能可以胜任审判者的角色。在法官为主、机器为辅的分工下,仍延续了人机对抗语境下的“法官完人”预设——法官在专业能力上有着机器无法比拟的优势,能够审慎负责地甄别机器辅助建议,并排除案外因素的相关影响。

^⑩ 参见郭春镇、黄思晗:《刑事司法人工智能信任及其构建》,载《吉林大学社会科学学报》2023年第2期,第26-29页。

^⑪ 参见左卫民:《关于法律人工智能在中国运用前景的若干思考》,载《清华法学》2018年第2期,第122页;雷磊:《司法人工智能能否实现司法公正?》,载《政法论丛》2022年第4期,第81页。

^⑫ 参见宋旭光:《论司法裁判的人工智能化及其限度》,载《比较法研究》2020年第5期,第92页。

^⑬ 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》第2、5、8-11条。

^⑭ 参见[德]卢曼:《信任:一个社会复杂性的简化机制》,瞿铁鹏、李强译,上海人民出版社2005年版,第50-61页。

将审判信任奠基于对办案法官的人格信任,体现了一种对司法审判的简化思维方式。司法过程被化简为以审判者为中心的决策过程,审判信任被还原为对审判者的人格信任。既然法官作为审判者是可信的,那么法官主导的人机协同审判也是可信的。只要法官最终负责作出判决,其与人工智能的具体互动似乎成为法官自主决定的事务。拒绝抑或采纳机器辅助意见,则完全由法官自行作出裁量。此时,人工智能退居于法官的法袍之后,不与当事人、律师等诉讼参与人直接接触。这种做法具有简化复杂性的价值,法院组织似乎就不必对外披露人机协作的复杂互动过程。

综上,在既有人机协同审判模式下,审判信任依赖于办案法官的信任背书。法官被视为公平正义的守门人,通过发挥主观能动性和自由裁量权,对如何采纳人工智能辅助建议行使最终决定权,以寻找案件裁判的最佳正解。

(三) 法官人格信任的理想愿景和现实差距

在与人工智能的比较中,法官作为人类智识和公平正义的代言人,被赋予了“完人”的理想形象。法官既精通“经验法则、逻辑规则和审判技巧”,又恪守“良知、敬畏之心和正义理念”,作为法治正义之代表向社会传达司法权威。^⑮法官被视为正义的化身,这既包含了对法官这一社会角色的期望,也反映了塑造法官权威性的实际需要。司法实践中,对优秀法官个人事迹的颂扬,对司法机关赢得社会信任具有重要意义。“勤勉奉献,坚持严格公正司法,全心全意为人民群众服务,努力满足人民群众的司法需求”,这是对我国法官的一种理想化的标杆形象。^⑯

然而,仅仅基于对法官人格信任的理想化愿景,并不足以建立起审判信任的坚固根基。现实中的法官并非完美无缺,而是具有人性弱点的“常人”。像普通人一样,法官也有知识、经验上的盲点,可能出于自利的动机而采取机会主义行为,在认知和决策上也会受到非理性因素的影响。^⑰将审判信任建立在对法官的信任之上,并不是绝对可靠的。一旦理念与现实之间发生“想象断裂”,将导致法官人格信任变得异常脆弱,引起人们对审判信任的态度瞬间逆转。^⑱我国智慧法院建设的目的之一是预防和纠正法官办案过程中的潜在偏差和不当行为,从而实现审判质效的提升。只不过,在人机对抗语境下,法官作为常人的局限性被理想的法官完人形象暂时遮蔽。

人机协同审判中,人类法官处于一种“人在回路”情境,需承担起评估、纠正机器辅助建议的责任。^⑲然而,鉴于人工智能的技术复杂性、算法黑箱化等问题,要对机器辅助建议进行甄别纠正并非易事。特别是在遇到酌情量刑、再犯风险评估等缺乏标准答案的

^⑮ 参见前注⑧,潘庸鲁文,第117页。

^⑯ 参见宁杰:《论公共传播中先进法官的形象塑造》,载《人民司法(应用)》2016年第16期,第68页。

^⑰ 参见前注⑩,郭春镇、黄思晗文,第27-29页。

^⑱ 参见陈洪杰:《道隐无名:司法信任的经验路径与象征媒介》,载《北大法律评论》(第22卷第2辑),北京大学出版社2021年版,第307-308页。

^⑲ See Harry Surden, *Artificial Intelligence and Law: An Overview*, Georgia State University Law Review, Vol.35:1305, p.1320-1321 (2019).

问题时,法官能否发现并纠正人工智能生成结果的潜在偏误存在疑问。^{②①}此外,法官在人机协同决策过程中还可能受到办案压力、认知盲区等非理性因素的影响。故此,不能想当然地认为法官在作出审判时可以不受到算法偏误的影响甚至误导。

此外,将审判信任奠基于人格信任,未能考虑到人工智能与传统机械工具的差异,忽略了人类法官与人工智能的复杂互动过程。机械工具不存在情感或偏见,亦不参与对“意义”的主观解释,只是通过非语言的形式直接提供客观事实的信息。^{②②}就传统计算机应用系统而言,尽管电子元器件取代了齿轮、杠杆等机械部件,但仍可将其视为一种完全由使用者操控的机械性辅助工具。然而,在新一轮人工智能科技浪潮下,机器算法正展现出前所未有的智能化水平,并逐步承担起一定的意义解读工作,因而与传统辅助工具之间存在本质差别。在人机协同审判中,人工智能并非仅仅基于法官指令来完成法律查询、检索排序等简单任务,而是已经深度介入对法律和事实问题的分析。因此,如果审判信任完全奠基于法官人格信任,无疑是将人工智能视为完全受法官控制的惰性工具,低估了人机协同审判的风险与挑战。

(四) 从人格信任到系统信任的视角拓展

随着社会复杂性的不断提升,信任的客体对象从具体个人逐渐延展到组织系统、算法系统等非人格化的抽象客体。基于被信任对象的差异性,信任可被细分为若干不同类型。其中,最基本的信任对象是能与主体进行面对面交往的他人。此外,信任对象还可拓展到社会角色、组织机构、技术系统乃至社会系统等更加抽象的客体。若将人际信任作为圆心向系统信任迈进,信任对象的具象程度逐渐减弱、抽象程度不断增强。所谓“人格信任—系统信任”的经典二分法,^{②③}是对前述信任谱系的一种简化表达。而本文所探讨的法官人格信任,既指向具体的法官个人,又与作为社会角色的法官群体紧密相连。法官本身就是一种容易唤起信任感的专业型公职角色,而对优秀典型的宣传又能进一步提升人们对法官职业角色的信任。^{②④}

人们对法官的人格信任,实际上是以司法系统的总体信任作为制度背景。人格信任建立于对他人的道德品质、专业能力的信赖,而对系统的信任则奠基于系统运作机制的正确性、合理性,更强调对抽象原则和技术知识的信赖。尽管存在差异,人格信任与系统信任之间并不是相互替代、非此即彼的关系,人格信任可以成为系统信任的一种表征机制。由于系统概念过于概括、抽象,人们容易基于感性经验而对人格化对象产生信任的心理纽带,这是个体在适应社会复杂性时发展出的简化机制。就审判活动而言,司法系统内部的运作细节对普通人而言显得过于陌生、抽象。相比之下,法官是司法系统与公众的“交汇口”,人们通过亲身经历、道听途说或媒体宣传等渠道,往往对法官有更强的熟悉感。在实践中,树立理想的法官形象也是法院赢得外部信任的重要方式。由于法官

^{②①} 参见李本:《美国司法实践中的人工智能:问题与挑战》,载《中国法律评论》2018年第2期,第56页。

^{②②} 参见〔法〕布鲁诺·拉图尔:《我们从未现代过》,刘鹏、安涅思译,上海文艺出版社2022年版,第49-50页。

^{②③} 参见前注^{②①},卢曼书,第62-79页。

^{②④} 参见〔波兰〕彼得·什托姆普卡:《信任:一种社会学理论》,程胜利译,中华书局2005年版,第57页。

吸引了公众的注意力,此时作为制度背景的司法系统则变得模糊起来。

对审判信任的分析,不应止步于人格信任,而应将视野扩大到对整个司法系统的理解和信任上来。从系统视角来观察,法官只是形成司法审判过程中的一个环节。法官的审理过程不能任凭主观恣意,而必须基于法律、事实和程序。正是出于对人格信任的审慎怀疑,司法系统才建立起一系列限制主观恣意的组织机制甚至技术性手段。故此,对审判信任的思考,应进一步扩展到对审判机关的系统信任之上。

三、审判机关制度信任的异化风险

人们对审判人员的人格信任,不能脱离于对审判机关的系统信任,故对审判信任的分析应从法官个体延展到作为整体的法院组织。法院借助诉讼程序、司法公开、司法责任制等一系列制度安排,保障司法审判的公平公正。故此,对法院组织的信任,最终奠基于制度信任之上。在人机协同审判模式下,法院组织的相关制度规则能否发挥预期作用,抑或可能存在失效乃至异化的风险,这是需要进一步探讨的问题。

(一) 审判机关系统信任的制度内核

法院作为行使国家审判权的审判机关,是较之法官而言更为抽象的组织系统。当法院作为信任对象时,其所对应的信任类型属于系统信任。社会大众未必了解法院内部的运作原理和具体细节,但仍可能形成对法院组织的信任。正如吉登斯所指出的,人们常常需要信任那些超出其理解能力的抽象系统。^{②④} 系统信任一定程度上促成了信任与理解的分离。普通人未必理解医学、法律或科技系统的复杂性,但仍能信赖医生、法官和工程师的专业判断。故此,当人们对法院组织有足够的信任时,尽管未必知晓其内部运作细节,但仍能对法院审理结果产生信赖和认可。

法院组织系统(以下简称“法院组织”)具有专家系统和科层组织的双重色彩。“专家系统”(expert system)是对专业知识与技术、业务专家进行组织和整合的系统机制,为人们在现代社会提供可靠的决策支持。^{②⑤} 法院通过组织和管理办案法官等司法人员来开展审判业务,处理法律纠纷、解释法律规则,从而构成一个典型的专家系统。同时,作为国家审判机关,法院亦具有科层组织的色彩。^{②⑥} 作为层级化的组织系统,法院通过规则、程序对其成员的职责和权限加以明确。法院还通过绩效评价、量刑指引、类案检索、监督惩戒等审判管理机制对法官审判工作进行规范、激励和约束。

对法院组织的信任,归根到底在于对其相关制度规则的信任。相较于人格信任,人们之所以会对抽象系统产生信任,主要原因不在于隶属于该系统之个体的品质或动机,

^{②④} See Anthony Giddens, *The Consequences of Modernity*, Polity Press, 1990, p.27-28.

^{②⑤} See Anthony Giddens, *Modernity and Self-Identity: Self and Society in the Late Modern Age*, Polity Press, 1991, p.18-19. 此处所称“专家系统”是一个社会学概念,并非模拟人类专家决策能力的计算机系统。

^{②⑥} 参见苏力:《审判管理与社会管理——法院如何有效回应“案多人少”?》,载《中国法学》2010年第6期,第186-187页;重庆市高级人民法院课题组:《审判管理制度转型研究》,载《中国法学》2014年第4期,第95-98页。

而在于对制度、规范、机制或原理的信任。^{②7}对组织系统的信任具有“去人格化”的特点,不能还原为对组织中个体成员信任。^{②8}法院不是法官个体的简单、机械地聚合,而是通过一种系统、有序的方式加以有机整合,从而形成一个整体性的抽象系统。法院通过机制设计、职业伦理、问责机制等一系列手段对法官个体的行为进行约束,从而保障审判活动在整体上是可信的。诚然,法院赢得公众信任的途径并不仅限于制度建设,还包括树立模范法官、普法宣传、媒体报道等其他方式。但不可否认,制度信任构成了法院组织系统的信任内核。尤其在法治社会背景下,法院是对其他社会主体进行监督和强制的“责任机构”(agencies of accountability),^{②9}对于守护契约精神、优化营商环境、建立诚信社会具有基础性的支撑意义。作为社会整体信任的基石,审判机关的系统信任必然要以法治精神为本色,以制度之治和规则之治为内核。因此,对审判机关的系统信任进行分析,就要深入到相关制度规则的设计原理和运作逻辑之中,考察其能否在人机协同审判模式下继续维系其制度信任。

(二) 人机协同审判“黑箱化”与司法责任制

随着人工智能技术在法院审判中的广泛应用,人机协同审判模式给法院组织的传统制度结构带来了深刻挑战。一方面,人工智能技术本身仍面临机器幻觉、可解释性差等问题;另一方面,人工智能对法院审判活动的介入,可能导致法院组织原本就不得不防范的潜在风险,如偏见歧视、机械司法、官僚司法等,进一步被加剧和放大。^{③0}故此,如果无法打消人们对前述风险的担忧,法院组织的制度信任将陷入危机之中。

法院制度信任所面临的核心挑战是,如何使外界信任人机协同审判这一具有“黑箱”色彩的决策过程。对黑箱的关注,主要滥觞于“算法黑箱”的相关研究。更广泛意义上,黑箱这一隐喻强调决策过程的不透明性,因而不仅限于计算机技术领域。司法审判活动中,合议庭评议、审委会讨论等环节并不对外示人,法官内心的心证过程亦只有其本人能直接感知。就此而言,传统司法过程本身带有一定程度的决策黑箱色彩。正是借助制度、机制和程序来约束审判过程中的决策黑箱,法院组织的系统信任才得以建立起来。随着人工智能介入法院审判,人机协同审判构成一个区别于传统审判的新型决策活动。虽然法官是最终的裁判者,但人工智能可以通过影响法官心证来间接影响审判结果。人机协同审判是“算法黑箱”与“人脑黑箱”相互叠加形成的复杂决策过程。算法黑箱具有一定的稳定性、连续性和普遍性,人脑黑箱则更具个别性、随机性和不断变化的特征。^{③1}当然,法官可通过释明权、心证开示等手段解释其决策过程,但真实的心证情况只有法官本人知晓。因此,如何约束法官与算法的互动过程,避免偏离亲历性、独立性、判

^{②7} See Anthony Giddens, *supra* note 24, 33-34.

^{②8} See Russell Hardin, *Trust in Government*, in V. Braithwaite & M. Levi eds., *Trust and Governance*, Russell Sage Foundation, 1998, p. 21-22.

^{②9} 关于“责任机构”概念的分析,参见前注^{②8},彼得·什托姆普卡书,第63-64页。

^{③0} 参见季卫东:《人工智能时代的司法权之变》,载《东方法学》2018年第1期,第128页;焦宝乾、赵岩:《人工智能对法官思维的影响》,载《求是学刊》2022年第4期,第122-123页。

^{③1} 参见马长山:《智慧社会背景下的“第四代人权”及其保障》,载《中国法学》2019年第5期,第6-7页。

断性等司法规律和原则,需要法院作出制度性的回应。

面对上述信任挑战,法院主要通过重申司法责任制和法官主体地位予以回应。司法责任制的内涵可概括为“让审理者裁判,由裁判者负责”,旨在强调人类法官的权责一致性。在人工智能时代,法院组织进一步强化了审判人员的主体地位和司法责任——“人工智能辅助结果仅可作为审判工作或审判监督管理的参考,确保司法裁判始终由审判人员作出,裁判职权始终由审判组织行使,司法责任最终由裁判者承担”^②。相比纯粹的法官人格信任,司法责任制借助问责机制对办案法官形成外部约束,督促办案法官对算法辅助意见中的偏误进行纠正。基于这一逻辑,实践中的人机协同审判呈现两大特点:一是“法官为主、机器为辅”的内部分工,二是“法官在明、机器在暗”的外部形态。关于法官与算法在个案中如何互动,目前在庭审公开、裁判文书等公开渠道中均无法觅得踪迹。就外观而言,人机协同审判与传统审判活动之间似乎没有变化。

相应地,法院组织对人机协同审判的制度安排,是将人工智能消隐于法官主体性背后,使司法责任制作为维持法院制度信任的逻辑支点。这一制度安排背后的潜在逻辑是,在巩固法官司法责任制的基础上,法院目前的制度结构足以吸收人机协同审判带来的风险与不确定性。既然人机协同审判的最终裁判者仍是审判人员,则约束审判人员的传统制度逻辑仍能继续奏效,所以法院组织的制度信任理应由此得以维系。然而,前述逻辑是否经得起推敲,仍需进一步结合人机协同审判的特点予以审视。

(三) 预测性审判与法院组织的制度异化

司法人工智能的算法逻辑,概括来说是基于裁判文书数据来预测和模拟司法裁判。目前,主流的人工智能模型大多基于神经网络和深度学习技术进行训练和预测。在训练数据方面,由于司法裁判文书的数据资源丰富,文本内容结构化、标准化程度较高,因而成为法律人工智能领域的主要数据来源。至于案件卷宗、合议庭讨论等数据,由于敏感程度高、质量良莠不齐等原因,并未得到广泛使用。可以说,司法人工智能本质上是一种概率预测模型,其通过分析裁判文书中的判决概率趋势,以生成辅助性审判建议。

司法人工智能具有的预测性审判特征,与强调亲历性、独立性、判断性的传统司法原则存在冲突。作为审判辅助工具,人工智能的核心使命是基于历史数据作出准确的预测,而不是对结果是否公平正义进行评价。这不禁使人联想到霍姆斯所描绘的司法预测论:不在乎公理原则或规则演绎,只对法院的判决进行预测。^③类似地,只要算法生成的结果被法官所采纳,就意味着预测是准确的。人工智能这种预测性审判的算法逻辑,难以实现基于情理法的利益衡量与价值判断,因而可能会产生算法偏见、机械司法等弊端。^④诚然,传统的司法审判活动也可能存在机械司法、司法官僚化等风险。经由机器算法的数据挖掘与模式识别,相关风险可能进一步扩散和恶化。在极端情况下,司法审判甚至可

^② 《最高人民法院关于规范和加强人工智能司法应用的意见》第5条。

^③ 参见[美]霍姆斯:《法律之道》,姚远译,载周赞主编:《厦门大学法律评论》(总第26辑),厦门大学出版社2015年版,第158-159页。

^④ 参见前注^①,雷磊文,第74-78页。

能彻底沦为例行公事的走过场,人们对司法审判的信任将遭到严重破坏。

在预测性审判的影响下,人类法官对司法人工智能的态度可能从“对抗”迈向“合谋”。所谓“人机合谋”,是人机协同审判的一种极端形态,人类法官对司法人工智能推送的建议一概予以采纳。发生“人机合谋”时,机器辅助建议的采纳率不断攀升,法官对相关建议的偏离率不断降低,法官审理结果与机器辅助建议将难分彼此。此时,法官在名义上仍是裁判者,事实上却可能在盲目接受算法推送的建议。实际上,“人机合谋”的风险并非耸人听闻。对司法人工智能而言,预测准确性是评估其性能的核心指标,算法的更新迭代会自然筛选出采纳率更高的模型。对办案法官而言,机器辅助建议提供了新的观察视点,使其能对法官群体的审判行为进行预测。由此,机器辅助建议可能在法官群体内形成类似“焦点博弈”的效果。^⑤法官个体通过调整自身意图或期望以匹配他人,最终使算法输出的结果成为群体的共识选择。

面对“人机合谋”的风险,法院既有制度对法官的激励约束可能会失效甚至异化。首先,采纳机器辅助建议可以大幅提升法官单位时间内的办案数量,在“案多人少”背景下完成更多办案指标。以文书辅助生成功能为例,人工智能可以在几秒钟内生成完成度颇高的判决书,点击“一键生成文书”无疑对法官极具吸引力。^⑥如果法官拒绝采纳机器辅助建议,就需要投入更多的时间和精力。其次,法官通过采纳人工智能辅助建议,还能降低被司法问责的风险。既有司法责任制背景下,“是否应责”的问题往往依赖法院在内部开展调查认定工作。法院内部问责时以类案裁判的“标准统一”为指针,以对标准的偏离为风险。^⑦目前,偏离度预警已融入法官的日常工作内容,并成为法院问责的重要线索。^⑧而随着人工智能预测日益准确,辅助建议将逐步成为标准答案,在无形中迫使办案法官向其靠拢。^⑨因此,从法官个体视角而言,依赖机器辅助建议进行判案,不但能缓解办案压力,还能避免因偏离预警而触发问责。

综上,法院组织尚未对人机协同审判作出完善的制度安排。现行制度未能充分考量预测性审判对法官激励约束的影响,无法有效纾解算法逻辑与法院科层制叠加后可能引发的制度失灵风险。在人机协同审判的变革中,法院组织的制度信任已经出现裂隙。

四、人工智能算法信任的局部补强

随着人工智能对社会生活的不断渗透,如何建立对人工智能技术本身的信任成为一项难题。作为一个垂直应用场景,人工智能司法应用同样受到准确性、可解释性、可问责性等算法信任问题的困扰。相应地,提升人工智能系统的可靠性,对审判信任的构建具有

^⑤ See Thomas C. Schelling, *The Strategy of Conflict*, Harvard University Press, 1980, p.57-58.

^⑥ 参见丁国锋:《公正路上, AI 和法官并肩作战》,载《法治日报》2024年5月29日,第4版。

^⑦ 参见最高人民法院《关于深化司法责任制综合配套改革的实施意见》(法发〔2020〕26号)第9条。

^⑧ 参见樊传明:《被敞视的法官:数字司法对审判权运行的影响》,载《法制与社会发展》2024年第3期,第137页。

^⑨ 参见前注⑪,雷磊文,第77页。

积极意义。然而,算法信任主要聚焦于对技术系统的信任,难以涵盖当事人、法官以及人工智能在审判活动中的复杂互动过程,故而只能局部实现对人机协同审判的信任补强。

(一) 人工智能算法信任的基本理念

在人机协同审判的语境下,算法信任即对法院所部署的人工智能审判辅助系统的信任。人工智能的技术初衷源于对人类智能的模拟,而随着相关技术的不断突破,关于人工智能的主体性探讨亦未曾停歇。^{④①}然而,对司法人工智能的算法信任本质上是一种系统信任,而非人格信任。我国各级法院所采用的人工智能审判辅助系统,是由多元化、异质化算法模块构成的复杂信息系统。因此,司法人工智能应被视为一种抽象的技术系统,而非拟人化的道德行动体。相应地,人工智能的算法信任是一种独特的系统信任。

人工智能系统的准确性、安全性和鲁棒性等技术指标,构成了算法信任的基本要求。值得注意的是,前述指标并非专门针对人工智能系统,而是在操作系统、数据库、高性能计算等传统信息系统中也同样适用。首先,准确性是人工智能赢得用户信任的基本门槛。如果系统频繁出现错误,用户显然难以对其产生信任。在司法实践中,使用准确性欠佳的机器算法不仅毫无意义,还可能干扰正常的决策流程。^{④②}其次,安全性是算法信任的另一重要维度。一旦发生数据泄露、隐私侵犯等安全风险事件,可能会彻底摧毁用户信任。在人工智能时代,随着数据规模、模型复杂度的急剧提升,安全问题变得更加严峻,任何疏漏都可能引发广泛的信任危机。最后,系统鲁棒性是指技术系统面对不确定性、干扰或异常情况时仍保持平稳运行的能力。人工智能在实际生产环境中肯定会遇到与训练环境有所差异的场景,其在这些未知情况下的表现也会对算法信任带来一定影响。由此可知,尽管人工智能在数据规模、模型复杂度等方面远超传统信息系统,但准确性、安全性和鲁棒性等技术指标依然是影响算法信任的重要因素。

除上述指标外,人工智能的可解释性与可问责性问题近年来备受瞩目。算法信任不仅关乎人工智能系统本身的技术可靠性,更与人们对技术的认知理解及熟悉程度密切相关。借由海量数据与复杂模型的深度融合,人工智能系统展现出令人惊叹的“智能涌现”能力,^{④③}同时引发了其可能超越人类理解、脱离人类控制的深切忧虑。通过对机器算法的可解释性与可问责性要求,可以在一定程度上实现对技术系统的透视与干预,从而增强人们对人工智能的信任感。^{④④}同时,借助可解释性与可问责性,能够构建针对机器算法的制度约束,提升人们对算法系统的控制力。^{④⑤}因此,可解释性与可问责性成为探讨人工智能算法信任问题的重要切入点。

^{④①} See Simon Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law*, Cambridge University Press, 2021, p.113-143.

^{④②} 参见前注^{④①},左卫民文,第124页。

^{④③} See Jason Wei et al., *Emergent Abilities of Large Language Models*, at <https://arxiv.org/abs/2206.07682>. (Last visited on Jan. 13, 2025).

^{④④} 参见苏宇:《算法规制的谱系》,载《中国法学》2020年第3期,第172-173页。

^{④⑤} 参见张欣:《从算法危机到算法信任:算法治理的多元方案和本土化路径》,载《华东政法大学学报》2019年第6期,第26-27页。

（二）人工智能算法的可解释性与可问责性

人工智能系统的可解释性和可问责性,直接影响着算法信任的预期建立与维系。建立信任的前提是形成对系统决策行为的特定预期。^{④⑤}可解释性的价值在于帮助人们理解算法系统并形成稳定的预期,而可问责性的意义在处理预期可能落空的情况时得到体现。具言之,当算法系统能对其运作过程提供清晰的解释时,外部用户就形成合理的预期,并逐步建立对输出结果的信赖。而当算法系统出现差错偏误时,可问责性有助于妥善处理、吸收由于预期落空而产生的不满。

1. 算法可解释性

算法可解释性是构建和维系人工智能算法信任的重要基础。可解释性强调人工智能系统能被人们所理解,使后者得以建立熟悉感和合理预期。算法规制的早期研究非常强调代码公开。例如,劳伦斯·莱斯格认为,既然在数字空间中“代码即是法律”,就应像公开法律文本一样公开代码。^{④⑥}然而,对人工智能系统而言,问题并非简单源于代码被刻意隐蔽,而是由于模型算法的运作逻辑和技术原理过于复杂。以 DeepSeek、LLaMA 等开源大模型为例,其源代码已对外公开,但对代码的静态分析仍难以揭示出算法在具体应用中的决策逻辑。再加上知识库、调优技术、提示工程等方面的差异,甚至同一模型在不同场景中的表现也可能大相径庭。算法可解释性并不要求代码公开,而是聚焦于对算法运作原理进行描述,使其能被人们所理解。^{④⑦}同时,算法可解释性有助于从事实维度回溯模型开发和机器决策过程,从而为价值判断、制度设计和责任判定提供依据。^{④⑧}

由于人工智能系统的高度复杂性,算法可解释性在技术实现上仍面临一定挑战。在人工智能的早期发展中,基于逻辑规则的专家知识模型曾风靡一时。这类系统具有可解释的知识库和推理规则,但自我学习和泛化能力较弱。相比之下,目前广泛流行的大语言模型展现出令人惊叹的性能,但如何解释其数以亿计的参数的运作原理,仍是实践中的一大难题。人工智能并非对人类智能的简单模仿,而是有其特殊的存储和计算逻辑架构。正如有学者指出的,可解释性与模型性能之间存在紧张关系,过度追求算法可解释性,有点像在强迫人工智能“人为地愚蠢到我们可以理解它是如何得出结论的”^{④⑨}。因此,试图以人类认知尺度来解释人工智能算法,其技术可行性上仍有待论证。

2. 算法可问责性

算法的可问责性是巩固算法信任的重要保障。如前所述,人工智能系统的可解释性与性能表现存在潜在冲突,加上人类思维在认知复杂性上的局限,可解释性相关技术的研究仍面临巨大挑战。因此,借助事后的结果问责以摆脱对技术细节的纠缠,能在一定

④⑤ 参见肖红军:《算法责任:理论证成、全景画像与治理范式》,载《管理世界》2022年第4期,第209页。

④⑥ See Lawrence Lessig, *Code: Version 2.0*, Basic Books, 2006, p.152-153.

④⑦ See Alejandro B. Arrieta et al., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*, Information Fusion, Vol.58:82, p.87-91 (2020).

④⑧ 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,第191页。

④⑨ [意]埃琳娜·埃斯波西托:《人工沟通与法:算法如何产生社会智能》,翁壮壮译,上海交通大学出版社2023年版,第25页。

程度上缓解对可解释性的需求。^{⑤⑩}通过采取严格责任、强制保险等制度规则,法律无需在技术泥淖挣扎就能实现事后问责。此外,可问责性还能用于推动可解释性的落实,敦促人工智能的系统研发者开发出可解释性更高的算法模型。而当研发者无法作出必要解释时,则可能需要承担相应法律责任。^{⑤⑪}

当人工智能系统出现问题或未达预期时,可问责性能够有效吸收风险和不确定性带来的负面影响。算法可问责性的要求是,确保算法系统是负责任的、安全的,并建立相应的审计、归责和补救等机制。^{⑤⑫}通过事后问责,人工智能系统的研发者和使用者被要求对其行为负责,从而促进机器算法的迭代与优化。以人机协同审判为例,在审判人员、算法系统、技术服务商的交互中,任何环节的问题都可能造成复杂的连锁反应,加剧责任认定和分配的难题。^{⑤⑬}因此,在算法较为复杂时,基于可问责性的规制方式能倒逼系统研发者采取相关措施,避免因系统设计、使用过程中的疏忽或过错而导致的不利后果。

(三) 算法信任对审判信任的局部补强

建立人工智能的算法信任机制,对人机协同审判的信任构建具有积极意义。准确性、可解释性、可问责性等评价指标构成“可信人工智能”的一般性要求,因而亦能适用于人机协同审判这一应用场景。《最高人民法院关于规范和加强人工智能司法应用的意见》第6条要求,司法人工智能从研发到应用全流程可解释、可验证,确保运行结果可预期、可追溯、可信赖,这与算法信任的相关理念不谋而合。

然而,算法信任主要着眼于人工智能系统本身,并非人机协同审判的互动过程,因而只能对审判信任发挥局部补强的作用。而一旦缺乏对人机协同过程的系统性剖析,则容易囿于主体与工具二元对立的思维定式,难以对法官与算法的具体互动形成约束。^{⑤⑭}实际上,人工智能对司法审判的介入是间接性的,需要以办案法官为媒介,通过影响法官的心证状态从而影响审理结果。至于法官究竟如何使用人工智能,则超出了算法信任本身的讨论范围。诸如法官如何评估、怎样取舍、是否采纳机器辅助意见,以及机器辅助建议是否需要对外进行披露,这些问题难以直接从算法信任的理论框架中找到答案。于是,人机协同审判在实践中存在一种“两层皮”现象:尽管人工智能的可解释性、可问责性等问题在法院组织内部颇受重视,但外部人仍无从知晓人工智能如何影响个案审理。

此外,过度强调对人工智能算法的信任,可能会对人格信任和制度信任产生冲击。从信任主体的视角看,法官、制度和算法作为潜在信任对象,相互间有一定功能等价性。在司法人工智能首次亮相时,由于算法信任尚未建立,人们对法院和法官的信任度较高,后者拒绝或偏离机器建议的阻力也较小。随着机器建议的采纳率逐步提高,人们对算法

^{⑤⑩} 参见沈伟伟:《算法透明原则的迷思——算法规制理论的批判》,载《环球法律评论》2019年第6期,第33-38页。

^{⑤⑪} 参见李训虎:《刑事司法人工智能的包容性规制》,载《中国社会科学》2021年第2期,第59页。

^{⑤⑫} See Joshua A. Kroll et al., *Accountable Algorithms*, University of Pennsylvania Law Review, Vol.165:633, p.636-642 (2017).

^{⑤⑬} 参见衣俊霖:《数字孪生时代的法律与问责——通过技术标准透视算法黑箱》,载《东方法学》2021年第4期,第83-84页。

^{⑤⑭} 参见郑戈:《人工智能与法律的未來再思考》,载《数字法治》2023年第3期,第78页。

的信任可能会逐步提升。如果社会上缺乏对法院、法官的充分信任,则很可能发生信任转移现象,寄希望于让看似更中立、客观的人工智能来作出裁判。而随着人们对“AI法官”的熟悉度不断提升,算法信任逐渐深入人心,审判信任可能会蜕变为算法信任。法院和法官一旦偏离人工智能辅助建议,将面临巨大非议,从而导致审判权的正常运行受到干扰。不难发现,前文所述的算法逻辑压过司法逻辑的风险再次出现。只不过,向算法靠拢的压力此时不来自法院组织而是来自社会大众。

综上,对算法信任的研究,触及了人工智能系统的可信性,而不再想当然地假定法官能纠正机器辅助建议的任何问题。然而,算法信任及其相关规制策略,仍处于人机分立、主客二元的静态分析框架内,未能充分考虑法官与机器、法院组织与算法系统的动态协同过程。而人机协同审判的信任难题,不仅在于算法模型的技术复杂性,更在于人机协同互动的沟通复杂性。尽管算法信任有助于审判信任的构建,其也只构成解决方案的一个局部。因此,对审判信任的探讨不能止步于算法信任。

五、基于沟通议论的整合式信任构建

无论是对法官人格、法院制度,还是对人工智能算法的信任,都仅是审判信任的局部环节,任何一种信任类型都无法单独构成人机协同审判的信任基石。关键在于,如何找到统合这些信任对象的纽带,从而在整体上构建起对人机协同审判的信任。要实现三种信任类型的整合,需聚焦于信任主体与信任对象的相互关系,通过法律沟通与议论,实现审判信任的动态构建。

(一) 信任构建的沟通互动原理

人机协同审判的信任构建,建立在多个信任对象相互支撑的整全性网络上,不存在单一的信任基础。在该信任网络中,人格信任、制度信任以及算法信任均发挥着重要价值。不过,三者各自亦只构成审判信任的一个局部,相互之间并不是简单的线性递进关系,而是循环指涉的网状关系。审判信任不能还原为单一信任类型,不存在“阿基米德支点”。或有观点认为,既然审判信任不存在单一基础,将其建立在“复合型”信任基础之上即可。^{⑤⑤} 难题恰恰在于,不同的信任类型如何统合为一个整体。在静态观察角度下,可以从审判信任中析出人格信任、制度信任以及算法信任等不同的信任类型。分别来看,亦可针对不同信任对象提出补强措施。然而,信任对象相互之间的裂隙仍会在相互指涉过程中不断激荡、放大,最终可能导致整个信任网络的崩溃。因此,审判信任的构建需要将人格信任、制度信任和算法信任整合为一个融贯的信任网络。

统合人格信任、制度信任和算法信任的关键,在于推动信任主体与信任对象的沟通互动,实现三种信任的反思性整合。信任存在于信任主体与信任对象的交互关系之中。^{⑤⑥}

^{⑤⑤} 参见赵杨:《人工智能时代的司法信任及其构建》,载《华东政法大学学报》2021年第4期,第76-79页。

^{⑤⑥} See Russell Hardin, *Trust and Trustworthiness*, Russell Sage Foundation, 2002, p.9-11.

信任的构建过程,不是静态、孤立、单调线性的,而是动态、整体、循环往复的。经由信任主体的审慎反思,原本相互割裂的人格信任、制度信任和算法信任相互调适与平衡,从而形成统合、融贯的信任(或不信任)评价。如果主体作出积极的信任判断,就意味着其在反思的意义上对诸信任对象实现了整合;而如果选择消极不信任,则意味着对局部环节有所疑虑,抑或认为信任对象相互之间存在鸿沟和断层。经由动态持续的沟通互动,主体能对各个局部作出信任或不信任的判断,并根据新的信息或经验来调整总体上的信任程度。

通过沟通与互动,信任主体得以获取信任对象的相关信息,从而对自身的信任态度进行反思和校准。在面对面的人际交往中,沟通互动对信任构建的价值几乎是不言自明的。“一个一般的规律性似乎表明紧密、亲近、熟悉敞开了相关信息的通道,也因此减少了操纵和欺骗的机会。”^{⑤7}而对法院组织、人工智能等抽象系统而言,他人难以从外部洞悉系统内部细节,所以可能因缺乏熟悉感而产生不信任。此时,信任主体与抽象系统之间的“信任校准”(trust calibration)就尤为重要。^{⑤8}开放、透明的信息获取渠道,有利于建立起更加坚固的信任基础。通过沟通互动,方能超越盲信或悲观,实现主体信任态度与对象可靠性相互对齐。

基于沟通互动的反思性整合机制,突出了信任主体在信任构建过程中的积极功能。在数字时代,随着社会复杂性的不断提升,信任对象的运作规律也持续演进、变化。相应地,人们对信息和知识进行持续更新以形成信任判断。信任的建立并非消极被动,而是具有积极、主动的面向。^{⑤9}这种积极的信任机制,与传统社会中“凝固的信任”形成鲜明对比,后者依赖于既定、静态的社会结构与规范。^{⑥0}换句话说,信任构建并非一劳永逸,而是一个具有协商性、开放性的持续过程。当然,在信任初步建立后,信任主体未必会时刻对信任对象进行审视和检查。但在必要的时候,信任主体唯有通过与信任对象的沟通互动,才能对信任基础是否坚实进行验证和评估。

综上,信任应当回归到信任主体与信任对象的沟通之中,实现人格信任、制度信任和算法信任的统合。一旦脱离了作为信任主体的当事人和社会大众,再多的专家背书、制度设计以及标准发布也难免陷入自说自话的境地。普通人对司法审判进行信任评估时,往往既不熟悉具体的承办法官,也不了解法院组织的内部复杂性,更不清楚人工智能系统的技术原理。尽管如此,人们仍可以从整体性、融贯性的视角进行综合评估,在深思熟虑的基础上形成对审判信任的总体判断。

^{⑤7} 前注^②,彼得·什托姆普卡书,第109页。

^{⑤8} See Bruno M. Henrique & E. Santos, *Trust in Artificial Intelligence: Literature Review and Main Path Analysis*, *Computers in Human Behavior: Artificial Humans*, Vol.2:1, p.9-11 (2024).

^{⑤9} 参见丁晓东:《基于信任的自动化决策:算法解释权的原理反思与制度重构》,载《中国法学》2022年第1期,第113-114页。

^{⑥0} 参见董才生:《论吉登斯的信任理论》,载《学习与探索》2010年第5期,第66页。

（二）审判程序的法律议论本质

沟通互动对信任构建具有重要价值,在司法审判领域尤其如此。由于审判程序具有公开、参与、中立、辩论等特点,法庭空间被视为认定事实、适用法律的最佳场合。从这个意义上说,审判程序归根到底是一个法律议论的过程。在司法审判中,沟通互动体现为具有话语博弈特点的法律议论。法官在进行审判时必须避免偏私,并充分听取双方意见。任何一方的主张无论本身多么正当有理,都要允许相对方意见的存在。唯有各方意见得到充分议论,才能避免对待决事项轻率得出结论的倾向。

司法审判活动中,当事人并非冷眼旁观,而是通过参与法律程序来影响审理结果。根据正当程序原则,与审判结果有利害关系的主体都有参与的权利,提出有利于自己的主张和证据,同时反驳对方的观点和证据。^⑥ 当事人的程序参与是审判结果得以正当化的重要理由:“在不断反驳和论证的话语技术竞争过程中,解决方案的多样性会逐步淘汰减少,直到最后找出一个大家都承认或者接受的正解,至少是唯一的判定答案。”^⑦ 因此,司法审判程序的核心价值在于保证法律议论充分开展。当事人经由法庭上的互动沟通与交锋论辩,直接参与判决结果的形成过程,并建立起对审理结果的信任。

对社会公众而言,其对司法审判的信任也在很大程度上源自审判程序的议论功能。并非每个人都有亲历诉讼的经验,因此,公众对审判信任的判断具有一种间接议论色彩。在某种程度上,公众将当事人视为自身的代表,通过观察后者在审判活动中的行动及其所受对待,形成对司法审判的信任判断。在最理想的情况下,审判信任能达到一种“胜败皆服”的状况——即便是承受不利结果的当事人,依然能在抽象公共人格意义上承认该结果符合普遍的公平正义。因此,确保法律沟通议论的公正性、透明性和参与性,不仅对当事人至关重要,也是赢得公众信任的关键。

人机协同审判信任困境的症结在于,司法审判中的沟通议论面临着转型困境。司法人工智能的引入,虽然提升了审判效率,但也可能导致法庭审判形式化,削弱司法过程的公开性和参与性。人工智能对司法审判活动的介入,在办案法官与人工智能之间形成了一个传统法庭之外的隔音空间。对法官如何使用算法缺乏明确的规则或标准,会导致人机协同审判的黑箱化问题。^⑧ 算法具体介入哪些裁判环节、法官对算法建议的采纳情况,这些信息均未对外予以披露。在此情况下,很难不使人产生怀疑:一切取决于法官与机器的私下互动,法庭上的面对面沟通已经名存实亡。

（三）人工智能角色功能的重新定位

目前,人工智能审判辅助系统主要服务于办案法官。通过将算法消隐于法官身后,法院保持了传统审判活动的外观。“法官在明、机器在暗”的互动模式,给审判信任带来一系列的消极影响。然而,倘若直接披露法官与人工智能的交互内容,这种穿透式方案虽然能约束法官对算法的过度依赖,却可能损害司法审判的效率和权威性。例如,法官

⑥ 参见[日]谷口安平:《程序的正义与诉讼》(增补本),王亚新、刘荣军译,中国政法大学出版社2002年版,第11-12页。

⑦ 季卫东:《法律议论的社会科学研究新范式》,载《中国法学》2015年第6期,第40页。

⑧ 参见前注②,李本文,第55页。

在庭审中利用算法查询法条或类案时,若直接披露相关结果,无疑会对庭审节奏造成干扰。在裁判文书的制作过程中,亦不宜披露机器生成的参考意见。否则,若裁判文书与参考建议一致,则使人怀疑法院是否沦为橡皮图章;若存在偏差,又难免引发对偏离原因的追问,削弱审判的终局性与权威性。因此,构建符合人机协同审判特点的沟通议论机制,是一项富有挑战性的任务。

为推动人机协同审判的沟通议论,应当对司法人工智能的角色功能进行重新定位,将其从内部办案工具拓展为外部沟通工具。目前,各级法院在评估司法人工智能的运用成效时,主要聚焦于办案效率的提升。智能类案检索、争点梳理、裁判辅助建议等应用大多以办案法官为主要服务群体。应当逐步将相关应用服务向当事人、社会公众等外部用户开放,使其具备与法官相近的技术能力,利用人工智能技术进行类案检索、法条检索及争点梳理等操作。当事人可对机器辅助建议的可靠性进行评估,及时发现问题并提交反馈,帮助预防和纠正人工智能系统的偏误。与此同时,法官仍可独立使用算法进行信息检索、量刑辅助、文书撰写等活动,而无须担心时刻处于外界审视之下。通过这种方式,人工智能不再是隐藏在法官背后的神秘黑箱,而是成为赋能各方开展沟通与讨论的普惠工具,同时也避免了对法院审判权的过度干预。

随着人工智能从幕后走向台前,基于沟通议论的信任构建过程将发挥作用,整合对法官、法院和算法的信任。当事人和社会公众通过与法官、法院及算法的互动,形成对人机协同审判的信任判断,从而纾解人格信任、制度信任和算法信任的困境。首先,针对法官人格信任的“想象断裂”问题,通过引入当事人的深度参与,公众可直接观察和评估法官的行为,增强对其专业判断力的信心。同时,法官的角色从单一决策者转变为沟通引导者,有助于重塑公众对法官的信任预期。其次,针对法院组织的制度信任问题,通过强化当事人对法律议论的实质性参与,可以推动法院在制度设计上更加透明与包容,降低办案法官与人工智能之间“人机合谋”的风险。最后,通过重新定位人工智能在司法程序中的角色,将其从效率工具拓展为沟通工具,可加强局部算法信任与整体审判信任的互动。当事人和公众能更直观地了解人工智能在审判场景中的功能定位和性能表现,并对其输出结果形成更清晰、准确的信任评价。

(四) 围绕庭审的人机协同法律议论

法庭审判对法律议论具有不可替代的意义,是各方沟通互动的中心舞台。平等对抗式的辩论原则促进了事实和法律的深入探讨,维护了司法审判的公平正义。因此,以下将结合庭审活动的相关阶段,说明如何在人机协同审判中推动沟通议论的具体展开。

在庭前准备阶段,司法人工智能可为当事人的庭审准备提供服务,并允许其对人工智能的偏误进行纠正。法院应向当事人提供法律法规、证据规则、类案检索等方面的人工智能辅助建议,促进当事人与法官之间的能力平等与信息对称。若当事人及其代理人发现机器辅助建议中的偏误,可及时提供纠正与反馈意见;甚至不妨鼓励当事人主动向法官提供类案检索报告等辅助材料,形成与机器辅助建议的观点竞争,促使法官实现兼听则明。更进一步,随着相关技术日益成熟,可以探索由人工智能直接组织双方进行证

据交换、争点梳理等事务性工作。在庭前会议结束后,人工智能可基于案件信息生成辅助调解建议,并梳理双方争议焦点。庭前阶段为当事人提供了与人工智能的深入互动的机会,使其得以发现问题、及时反馈,并能够为后续庭审活动中的沟通议论预做准备。

在庭审进行过程中,人工智能应实时梳理并展示诉讼双方的争议焦点,促进法律议论的实质化开展。目前,语音识别技术已经实现了对法庭辩论的如实记录。在此基础上,人工智能技术可以对法庭争议焦点进行动态梳理、归类和可视化展示。通过对诉讼请求、要件事实、证明反驳等论证因素的协议化和代码化,揭示“原告与被告之间争论点和妥协点”的话语互动,促进了双方你来我往的对话沟通。^④庭审结束后,将机器生成的争点清单连同庭审笔录一并提供给当事人,由后者进行查阅。当事人发现有遗漏或差错的,可借此机会请求补正。值得注意的是,人类法官在法庭审判中仍具有主导意义。法官应综合运用释明义务、心证开示等沟通手段,在反思机器建议对心证影响的基础上,及时向当事人进行提示,保障法律议论的充分展开。这不仅有助于保障庭审实质化、避免裁判突袭,也为法官建立人格信任提供了重要契机。

在判决书中,法官应当围绕案件争点释法说理,并对与争点结论密切相关的机器辅助建议进行披露。此时并非披露机器算法的运作逻辑,而是对其输出结果进行理由说明。例如,若法官裁判时将机器所荐类案作为对争点得出结论的重要参考,则应对案例的相关性进行说明。对于涉及数值计算的内容,如金额或刑期,法官应当披露人工智能所提供的基准、公式、参数等相关要素。通过在裁判文书中对机器辅助建议进行披露和解释,可以增强判决的透明度和可信度,允许社会大众对相关信息进行检验和甄别。因此,裁判文书是连接法院判决与公众理解、法律进步和社会反思之间的桥梁,对人机协同审判赢得公众的信任和认可具有重要意义。

综上所述,当事人与社会公众通过与法官、法院及算法的互动,形成对人机协同审判的信任判断,同时促进法官人格信任、法院制度信任与人工智能算法信任的完善。办案法官仍是人机协同审判的主导者,在庭审指挥、心证开示、释法说理等活动中发挥核心作用。法院组织通过对外提供人工智能服务、保障当事人程序参与,巩固其制度信任基础。人工智能作为促进法律议论的沟通工具,从幕后走向台前,由当事人对人工智能进行监督与反馈,推动其在准确性、公正性与可解释性等方面的持续优化。

六、结 语

在人工智能时代,审判信任的构建和维系显得尤为重要。信任存在于信任主体与信任对象的意向性关系之中,难免受到信任主体的个体知识和生活经验之影响,从而呈现出复杂多变、因人而异的情形。法学研究者比较了解法律系统的内部运作,既可能建立起更加坚固的信任,也可能认识到更多风险和不确定性而难以建立足够的信任。相比之

^④ 参见季卫东:《人工智能时代的法律议论》,载《法学研究》2019年第6期,第36-41页。

下,社会大众可能出于朴素的信任感而对司法审判产生信赖,或基于过往经验、道听途说以及社会舆论形成对审判的总体态度,而不会大费周章地辨析信任到底奠基于何处。本文对人格信任、制度信任、算法信任进行了批判性分析,揭示出审判信任的复杂性和多样性,并在此基础上试图寻找信任得以建立和维系的坚实根基。从短期而言,这些探讨恐怕未必能直接影响大众对人机协同审判的信任态度。但从长期来看,信任的构建本就是一个动态、互动的渐进过程。加强社会大众、法院系统、算法系统之间的沟通议论,可以形成一个允许信任主体亲身经历、仔细验证、审慎反思的信任构建机制。诚然,这个过程无法一蹴而就,但在此基础上所形成的审判信任,一旦建立起来便不会轻易动摇。

Abstract: With the rise of Artificial Intelligence (AI) in smart courts, judicial trials have been gradually shifting toward a human-machine collaborative model, where ‘the judge leads and the machine assists’. Currently, judicial AI remains temporarily hidden behind judges, with trial trust being replaced by personal trust in judges, systemic trust in courts, or algorithmic trust in AI. Personal trust, however, assumes judges are ‘perfect individuals’, ignoring their reality as ‘ordinary people’. While courts use judicial responsibility mechanisms to regulate the interaction between judges and algorithms, the predictive nature of algorithms risks the uncritical adoption of machine suggestions, leading to the potential risk of ‘human-machine collusion’. Although algorithmic trust serves as a partial remedy for trust deficiencies, it struggles to regulate the process of judge-algorithm collaboration due to the ‘subject-tool’ dichotomy. It is essential to conduct a reflective integration of personality, institutional, and algorithmic trust based on communicative interactions, redefine judicial AI as both an internal case-handling tool and an external communication tool, and fully promote the legal argumentation of human-machine collaboration, thereby laying a stronger foundation of trust for human-machine collaborative adjudication.

(责任编辑: 赵毅宇)