财经法学 No. 2，2023
pp. 3 - 20

人工智能司法的可解释性困境及其纾解

周　媛　张晓君*

内容提要：加强对人工智能司法发展及风险的研究是时代课题，其中人工智能司法的可解释性困境尤为关键。人工智能司法可解释性指的是司法决策或行为的可理解与透明性，涉及基础数据、目标任务、算法模型以及人的认知这四类关键要素。不可解释困境主要是由数据失效、算法黑箱、智能技术局限、决策程序和价值缺失等因素所致。但是，人工智能司法的不可解释困境其实是一个伪命题，可解释性具备认知层面和制度层面两方面基础。纾解困境的具体策略包括：构建司法信息公开共享制度，提高有用数据的甄别与利用效率；从软硬法结合视角建构司法系统的运行标准与制度规则；从全过程视角强化主体之间的协同治理；通过指导性案例和司法解释赋权法官的司法解释空间，提高法律解释技术；强化交叉学科人才建设，提高对人工智能司法决策模型的引领；发挥法官的自律与能动性，实现司法智能决策的人机协同。未来，不仅需要把握司法价值与技术理性的平衡，还需考虑人工智能对司法的差异化介入，推动人工智能司法战略目标实现。

关键词：人工智能　司法　算法　可解释性　协同治理

一、问题的提出

党的二十大报告明确提出："构建新一代信息技术、人工智能、生物技术、新能源、新材料、高端装备、绿色环保等一批新的增长引擎。"自 2015 年起，人工智能与司法工作深度融合发展战略上升为国家战略，各地纷纷开启智慧法院建设步伐。〔1〕从智慧司法 1.0 到 4.0，人工智能司法已成为一种现实，深刻地改变着传统法院的组织能力与管理结构，冲击着诉讼架构和程

* 周媛，上海交通大学凯原法学院博士研究生；张晓君，西南政法大学国际法学院教授。
本文为国家社会科学基金项目"城市更新中促进绿色建筑发展法律机制研究"（21BFX136）的阶段性成果。

〔1〕 如上海刑事案件智能审判系统、北京"睿法官"审判辅助系统、河北"智审"审判系统和浙江"金融智慧庭审平台"等。参见聂友伦：《人工智能司法的三重矛盾》，载《浙江工商大学学报》2022 年第 2 期。

序机制，重塑法律人的理念、情感、行为乃至结果模式，甚至影响整个司法权力在国家权力架构中的定位。〔2〕但对于智慧法院建设，学界呈现两种分歧立场：一种观点认为，从司法本质看人工智能司法具有主体正当性，从司法裁判的手段看智能裁判具有逻辑正当性，从司法过程看人工智能司法具有程序正当性，从司法结果看智能司法具有结果正当性，因此，人工智能司法的整体正当性充足。〔3〕另一种观点认为，人工智能司法的运用程度有限，只能作为一种实现司法正义的辅助手段，不能排斥法官的心证和裁量，这是其运用所应遵守的基本原则。〔4〕然而，法律是一种风险控制机制，习近平总书记在中国共产党第十九届中央政治局第九次集体学习时强调："要加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全，确保人工智能安全、可靠、可控。要整合多学科力量，加强人工智能相关法律、伦理、社会问题研究，建立健全保障人工智能健康发展的法律法规、制度体系、伦理道德。"〔5〕法律是风险识别和控制的主要手段。法律是一种社会建构。因此，对法律的合适态度，应该是审慎而非颂扬（celebration）。〔6〕法律的保守主义立场使得我们更需要保持一种批判主义研究进路。

目前学界对人工智能司法的批判性研究成果大致可归为以下几个方面：一是人工智能司法产品运用过程中外部技术环境的限制，最为典型的是作为智能司法决策基础的司法数据样本存在"伪充分性"，〔7〕还包括法律语言与计算机语言的隔阂等；二是人工智能内在的技术困境，典型的是算法歧视、算法黑洞与算法霸权；三是人工智能的伦理与价值困境，体现在人工智能无法对司法正义作出实质性权衡，也无法复制法律人的"情怀"和"匠心"；〔8〕四是人工智能受到司法环境的制约，如国家的政策性、政治性因素，地域因素，习惯规则以及法律体系的差异等。作为一项智能推理与决策技术，无论是人工智能司法数据的"伪充分性"、算法黑洞困境，还是人工智能的伦理与价值困境，都指向人类如何正确使用人工智能，而至于人工智能又如何满足人类对司法正义与价值目标的追求，透明性、可解释性及由此产生的可信赖性成了解决问题的关键。2019 年 4 月，欧盟委员会发布的《人工智能道德准则》提出了值得信赖的透明性规则；2021 年 1 月，欧洲议会和理事会发布的《关于人工智能的统一规则（人工智能法）并修正某些联合立法行为》同样对人工智能的透明性和可理解性进行了着重强调。〔9〕2021 年 9 月，我国国家新一代人工智能治理专业委员会发布了《新一代人工智能伦理规范》，明确提出人工智能发展需遵守"确保可控可信、强化责任担当"等六项基本伦理要求，基于可解释性才能实现验证、审核、预测与信赖。事实上，我国以人民为中心的司法审判工作本质是一种"回应型司法"或"纠纷解决型司法"，〔10〕回应型的内

〔2〕 参见徐骏：《智慧法院的法理审思》，载《法学》2017 年第 3 期。

〔3〕 参见彭中礼：《司法裁判人工智能化的正当性》，载《政法论丛》2021 年第 5 期。

〔4〕 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。

〔5〕 习近平：《加强领导做好规划明确任务夯实基础 推动我国新一代人工智能健康发展》，载《人民日报》2018 年 11 月 1 日，第 01 版。

〔6〕 参见〔英〕哈特：《法律的概念》（第 3 版），许家馨、李冠宜译，法律出版社 2018 年版，第 1 页。

〔7〕 参见前引〔1〕，聂友伦文。

〔8〕 参见马长山：《司法人工智能的重塑效应及其限度》，载《法学研究》2020 年第 4 期。

〔9〕 参见刘艳红：《人工智能的可解释性与 AI 的法律责任问题研究》，载《法制与社会发展》2022 年第 1 期。

〔10〕 参见〔美〕米尔伊安·R. 达玛什卡：《司法和国家权力的多种面孔——比较视野中的法律程序》（修订版），郑戈译，中国政法大学出版社 2015 年版，第 14－15 页。

在向度正是追求司法活动的透明性和可解释性。例如，行政诉讼领域的“行政争议的实质性解决”，法理逻辑即是通过充足的沟通与恰当的交流平台，让诉讼活动评价回归当事人的“体验感”，提高当事人的司法获得感，从而增强司法判决的可接受性。行政裁判的可解释性正是实质性解决行政争议的前提条件。

虽然近年来对于人工智能的可解释性的研究逐渐升温，但事实上人们对人工智能“黑洞”的内在原因及破解路径仍然疑问重重，就像2017年AlphaGo如何战胜了两位世界围棋冠军，赛后柯洁坦言其策略是那么让人惊诧。国内法学界聚焦人工智能司法的可解释性这类子领域的有力研究成果并不多见。〔11〕因此，下文从人工智能司法可解释性困境的具体表现入手，阐释人工智能可解释性的具体内涵，探寻可解释性困境的具体方面及其形成的机理，进而对可解释性的主客观基础作出研判，得出纾解人工智能司法可解释性困境的有效之道。

二、可解释性界定与人工智能司法可解释性困境

概念是逻辑研究的起点。对解释性的界定成为本文研究的起点。由于理解活动“具有双重的主观性：理解对象的主观性和自身活动的主观性”〔12〕，在对研究概念和对象进行界定和选取时要尽可能做到多维度、多视角。

（一）解释与可解释性：多学科融合视角

“解释”在汉语中包含两层词义：分析阐明和说明含义、原因、理由等。但在科学哲学领域，“解释”一词的词义往往从本体含义转移至语境功能层面，关联到解释主体和对象之间的逻辑关系。例如，亚里士多德就曾通过物理学的观察提炼出“四因说”，得出解释其实就是对事物（解释项）与事物（被解释项）之间“为什么”产生、发展、变化、消亡等一系列动因的说明。学者亨普尔（Carl G. Hempel）和奥本海姆（Poul Oppenheim）在《解释的逻辑研究》一书中进一步明确解释的“阐释”作用，即解释是对解释项与被解释项之间关系与逻辑的重构，或达到论证某种关联的目的。〔13〕由此，这种解释功能也被定义为对解释项与被解释项因果关系的挖掘，如美国哲学家刘易斯（Clarence Irving Lewis）直接将解释等同于因果关系的说明，“解释一个事件就是提供一些关于其因果历史的信息。在解释的行为中，一个拥有一些关于某个事件的因果历史的信息（我称之为解释性信息）的人试图把它传达给其他人”〔14〕。但是，因果关系并非事物关联性的全部，并非所有的因果关系都存在唯一的解释形式，由此，解释的路径和方法具有多元性，特别是科学哲学领域不同的逻辑将引导出不同的可解释性方法。例如，在知识图谱的推理方法中，至少存在“符号主义”“行为主义”“连接主义”“新型混合”四种可解释知识推理类型，不同的

〔11〕 相关文献参见前引〔9〕，刘艳红文；姚叶：《人工智能算法的不可解释性：风险、原因、纾解——兼论我国“举报人免责制度”的具体建构（英文）》，载《科技与法律（中英文）》2022年第3期；苏宇：《优化算法可解释性及透明度义务之诠释与展开》，载《法律科学（西北政法大学学报）》2022年第1期等。

〔12〕〔德〕马克斯·韦伯：《社会科学方法论》，韩水法、莫茜译，商务印书馆2020年版，序言第17页。

〔13〕 See Carl G. Hempel & Paul Oppenheim, Studies in the Logic of Explanation, 15 *Philosophy of Science* 135 (1948).

〔14〕 转引自闫坤如：《可解释人工智能：本源、进路与实践》，载《探索与争鸣》2022年第8期，第107页。

推理类型又对应着不同的推理方法。〔15〕这就需要以一种多学科融合的角度来理解人工智能领域的解释。

回到人工智能领域，人工智能领域的解释本质是指人工智能的可理解性或透明性，亦即人工智能的“演绎法则”能够被人所认识、领悟。从亚里士多德的“四因说”出发，人工智能的可解释性作为语境词汇，至少会涉及基础数据、目标任务、算法模型以及人的认知这四类关键要素。其一，数据是人工智能得以存在的前提，人工智能本质即是借用计算机对大数据强大的搜索统计、计量分析、深度学习等功能，而形成的一种计算认知，足够多、足够好与足够真实的大数据是人工智能的必要条件。但采集哪些数据不采集哪些数据，对于人工智能而言似乎并不是一个不言自明的清晰逻辑，它有待于目标任务的明确设定。其二，目标任务在逻辑上是为达到某种效果或完成某种行为，但如何将这种效果或行为设定转换为计算机语言以完成人类语言向数智语言的转变，需要抓住的关键是两类“语言”之间的联结点在哪里。一般而言，事物的规律性是两者之间的联结点，因此可以通过足够优质的大数据进行因果逻辑抑或形式逻辑的推理，建立两者之间的确定性或概率性的关联，形成一种以事物的规律性和逻辑自洽为中心的算法模式。其三，算法模型是数据充足、目标明确的基础上所刻画的一种客观逻辑结果，它也是人工智能可解释性的关键所在，按照可解释性的程度大致可划分出三类算法模型：〔16〕（1）参数模糊型，主要是指由于任务的复杂性与数据的繁杂性，算法模型并非设定一个单一的清晰的范围值或确定的任务目标，只是在一个可能的概率值范围内搜索更多的可能有用的数据，从而完成相对模糊的逻辑关联运算，如深度学习模型；（2）参数明确型，主要是指任务较为明确、逻辑较为清晰的目标运算，这类运算通常在给定的范围值内，搜集较大关联性的数据，从而得出较为公认或公式化的算法模型，如统计学习模式；（3）参数外显型，相比于明确型，这类算法本身具有透明性，也可称为“白盒模型”，需要提取哪些数据以及目标设定都十分明确，得出的结果也自然较为单一，通常遵循条件式因果推理规律，如专家知识模型。其四，人的认知是这四类关键要素的决定因素，相对于算法模型而言，人的认知与意志更具主观性，主观性会加剧对人工智能决策过程的理解差异，不同群体由于知识范围的差异对算法决策的理解能力自当不同，如何弥合主观与客观之间的鸿沟成为打开算法“黑盒”的金钥匙。

人工智能融入司法，相应的可解释性问题自然而然转移到对司法基础数据、司法目标任务、司法决策算法模型以及法官的认知这四类关键要素的理解上。首先，司法数据从内容和对象来看，是人民法院“在司法工作中形成的审判流程、执行信息、法律文书、庭审活动信息、司法政务、司法人事、外部协查等数据的总和及其关联关系”〔17〕；但从结果来看，它主要指向司法裁判文书以及相应的裁判技术。司法数据的存在是人工智能介入的基础，司法数据“质量”决定人工智能司法决策的真实性、可靠性，若是司法数据偏差自然会引发可理解性困境。其次，司法目标

〔15〕 参见夏毅、兰明敬、陈晓慧、罗军勇、周刚、何鹏：《可解释的知识图谱推理方法综述》，载《网络与信息安全学报》2022 年第 5 期。

〔16〕 参见刘桐、顾小清：《走向可解释性：打开教育中人工智能的“黑盒”》，载《中国电化教育》2022 年第 5 期。

〔17〕 孙晓勇：《司法大数据在中国法院的应用与前景展望》，载《中国法学》2021 年第 4 期，第 124 页。

任务是司法裁判要达成的政治效果和社会效果，我国司法的目标是以人民为中心，惩戒犯罪、化解社会矛盾，实现公平公正高效的司法服务。然而这一目标决定人工智能司法的行为逻辑或演绎规则必须纳入利益平衡与价值考量，而这正是计算机语言难以充分解释的又一困境之所在。再次，司法裁判的过程是对案件事实与证据、法律规则要素以及法官裁量标准的综合性判断，随着事实与证据的量化、法律规则以及法官裁量的标准化，司法裁判逐渐走向自动化，司法决策算法模型是自动化的一种表现，但事实与证据的模糊性、法律规则的滞后性以及法官裁量的主观性都可能导致算法决策模型难以形式化。事实上，目前的人工智能司法模型多体现为较为简单的统计学习型，如何进行智能司法深度学习是一大理论难题。最后，法官对法律的理解以及价值观的培育是司法目标能够达成的关键，统一法官的思维模式可保证司法目标的一致性，但在具体的案件类型、地区差异以及场景化司法运行过程中，法官基于实质平等的法律解释与续造所发挥的主观能动性恰是实现司法公平、公正的重要前提，而这增加了人工智能司法技术标准的设计难度，自然增加了不可解释性空间。司法这四个基础要素导入人工智能领域在实践中衍生出以下可解释性困境具体形态。

（二）人工智能司法可解释性困境的具体形态

第一，司法数据“低劣”引发人工智能决策失真。马云曾说，在21世纪数据好比支撑社会经济发展的“石油”。“司法大数据与人工智能技术的实质是建立了一种基于海量数据挖掘的认知范式，数据具有绝对的前置性。”〔18〕以数据为中心的司法智能是从相关的类案情节中提取判决规律的一种非理论预设的认知技术。足够多、足够优质的类案数据既是司法公正决策的前提，更是获得人们理解的基础。仅凭单个或特殊案例所总结的判决结论往往不具有较强的说服力，“坏的”“低劣”的数据也会导致人们对司法决策理解的偏差。从司法样本数据来看，“不充分”和“低劣”数据主要表现在：一是司法数据本身存在主观性，哪些数据可以公开、哪些不能公开，已经受到人为因素的左右，基于主观筛选和缺失样本所进行的司法决策很可能会背离实践真实样态，导致人们对司法决策能力和效果的误解。例如，在行政协议纠纷中涉及政府利益和商业秘密的案件一般不被公开，这些案件往往最能体现行政的本质；在刑事案件中，一般轻微伤或自诉案件可能按撤诉或协商处理，一旦撤诉或协商处理，相关定罪样本或要素就无法有效得以统计，计算机自然不能进行深度学习。二是司法数据分类或标记的科学性存疑。计算机是靠代码进行识别的，存在相应的代码才能进行统计或归纳，但对司法案件的代码分类或标记因素会因法官或研究人员的理解差异而参差不齐。例如，刑事判决书中，被告人的犯罪动机、犯罪环境等因素往往很难得到类型化标记；在刑事被告人赔偿案件中，赔偿数额及相关酌定情节也通常属于不被标记的范畴。因此，“这些未被标记的因素便会游离于所得数据之外，继而造成数据充分性问题，导致预测模型失准”〔19〕。三是司法数据的变换性使得精确性不足。在主观性之外，数据所依赖的环境同样重要。换言之，数据样本的真实性与数据所产生的特定环境密切相关，“人们在特定环境中分

〔18〕 王禄生：《司法大数据与人工智能技术应用的风险及伦理规制》，载《法商研究》2019年第2期，第102页。

〔19〕 前引〔1〕，聂友伦文，第68页。

析数据并将意义赋予了数据”〔20〕才能消除样本的分歧增进标注代码的可理解性。然而，也会存在司法异地差异、人为差异等因素导致的数据非真实性或系统性错误，这就会引发“错误的前提导致错误的结论”〔21〕(garbage in，garbage out)，加深人们对人工智能司法决策的误解。

第二，人工智能算法黑箱冲击司法透明，导致司法不公。算法是第二类关键性要素，但算法黑箱问题一直是困扰理论与实务界的头号难题。算法的不透明性大致存在三种情形：一是因国家秘密、商业秘密等保密之需要而形成的不透明，二是因技术的不成熟而引发的不透明，三是人工智能算法内部的复杂性而产生的不透明。〔22〕不论何种情况，由于价值平衡、技术有限以及民众知识差异等因素的制约，算法公开的程度都十分有限。这种限度不仅体现在由计算机技术人员和法律专业人员组成的审查机构审查的广度和深度有限，还体现在审查结果要尽量避免损害私人权利以及抑制新技术创新的负面效应。实践中，这三类不透明性往往交错在一起，加剧算法的不透明性。例如，目前概率建模下司法要素被压缩为几个方面，而采取启发式算法系统可模拟法官的思维，但这种思维决策过程是如何运作的往往无法被“追溯与验证”。〔23〕腾讯研究院也明确表示：“在 AI 深度学习模型的输入数据和输出结果之间，存在着人们无法洞悉的‘隐层’，深埋于这些结构底下的零碎数据和模型参数，蕴含着大量对人类而言都难以理解的代码和数值，这使得 AI 的工作原理难以解释”。〔24〕人工智能算法黑箱易引发的可解释性问题主要是，司法数据的关联性错误、法律适用的歧视性加深、司法决策的结果难以预测以及危及司法的公信力等。在司法关联性数据领域，由于缺乏足够的因果逻辑可能使得看似存在正负增长关联的事件之间，其实是一种虚假关联关系，从而可能错误地揭示司法规律。而由于无法看清司法决策规则，也不能参与其中，只能被动接受决策结果，往往可能导致算法偏见无法纠正而形成“滚雪球”效应，如美国黑人的犯罪率更高。同时，在无法理解算法模型的基础上，也可能导致结果的不可控性，特别是针对实质性的同类案件，某些外在参数不同竟然导致预测结果大相径庭。这种难以接受的算法偏见与歧视，自然也会影响司法的公信力。

第三，技术理性对司法理性的侵蚀导致内在隔阂。相对于算法黑箱而言，技术的有限性侧重于阐释目前人工智能应用于司法领域的技术瓶颈和固有缺陷。在类型层面，人工智能存在“弱”“强”“超”之级别区分，毫无疑问，目前人工智能司法只停留在“弱”人工智能阶段，其司法智能化程度并不高。由于人工智能是一种技术理性，司法决策是一种司法理性，司法理性“更多乃是依靠司法工作人员的认知、心性、德行并结合案件发生的客观现实环境”〔25〕，技术理性偏重于标准化与程式化的规则，因而，技术理性与司法理性之间存在一种天然的隔阂。加之法律语言与计算机语言的差异，法律与技术之间的融合存在较大的跨界障碍。从世界各国司法实践情况来

〔20〕〔德〕罗纳德·巴赫曼、吉多·肯珀、托马斯·格尔策：《大数据时代下半场：数据治理、驱动与变现》，刘志则、刘源译，北京联合出版公司 2017 年版，第 205 页。

〔21〕〔英〕维克托·迈尔、肯尼思·库克耶：《大数据时代》，盛杨燕、周涛译，浙江人民出版社 2013 年版，第 211 页。

〔22〕See Jenna Burrell，How the Machine “Thinks”：Understanding Opacity in Machine Learning Algorithms，3 *Social Science Electronic Publishing* 1 (2015).

〔23〕参见马靖云：《智慧司法的难题及其破解》，载《华东政法大学学报》2019 年第 4 期。

〔24〕腾讯研究院等：《可解释 AI 发展报告 2022：打开算法黑箱的理念与实践》，腾讯研究院 2022 年发展研究报告，第 2 页。

〔25〕陈灵峰：《司法人工智能的技术效应与应用边界》，载《求索》2021 年第 6 期，第 186 页。

看，人工智能系统主要运用于数据管理平台建设、类案检索和推送、证据审查和抽检等审判辅助层面，如美国联邦法院的“案件管理和电子案件档案系统”（CM/ECF）、欧洲的 ERP 系统及职位分配管理系统软件（OUT ILGREF）、新加坡的社区司法和裁判系统（CJTS）、韩国的电子案件归档系统（ECFS）等。〔26〕以辅助审判为主的“弱”人工智能司法意味着其决策本身的准确性与应用深度有限，“从某种意义上分析，这种应用程度及准确性缺陷导致的技术保守态度与解释困境是其内部机理不可解释性的重要原因”〔27〕。换言之，越是成熟的越具有实践经历的算法模型，随着时间的推移相应的黑箱问题越能迎刃而解。事实上，在人工智能算法模型建设初期，因计算机技术人员的水平有限、精力以及项目投入不足，经常导致算法模型的粗糙和固有缺陷，而引发不可解释性困境。

第四，逻辑运算对司法正当程序的冲击降低主体认同。法律程序很大程度上是一种交流和沟通装置，同时也是进行个人权益分配、提高商谈效能、保障人们尊严和维护自由主义传统的适当机制。〔28〕在司法领域，正当程序被视为“看得见的正义”的代名词，其首要贡献在于强调司法主体之间的“交往”，即“旨在通过使用符号（包括前符号、符号和元符号），来协调大家的行为和举止，以求得沟通和共识”〔29〕。人工智能司法决策程序的不透明直接导致这种“交往”受阻。在司法自动化模式下，由原告、被告、当事人、法院等多方主体参与的交互式辩论法庭模式，变成输入关键词得出结果的单一形式逻辑运算过程，后疫情时代催生的线上庭审模式一定程度上加剧了司法决策程序的形式化与非透明化。“这样的司法操作模式不仅会严重解构法官的主体性价值，削弱司法进行动态社会整合的内在张力，更会导致法律系统像停止生长的珊瑚礁那样，变成一堆毫无生机的死化石。”〔30〕更重要的是，缺失交往性的人工智能司法决策无法达致哲学层面“他者”向“自我”的身份更新与主体认同，因而就无法得到人们发自内心的司法信仰和精神服从。

第五，人工智能无法答责，引发司法归责模糊。“法律责任的本质是答责，不具有可解释性的人工智能不能自我答责，因此，其无法承担法律责任。”〔31〕责任承担问题又与人工智能的主体地位直接挂钩。意志自由是自我应责的前提，只有明晰责任后果、具有答责能力的人才是自由的主体。而在责任的功能层面上，责任往往与伦理道德一起共同发挥惩罚、修复或预防作用。在 H. D. 刘易斯看来，“责任……仅仅意味着做一个有道德的人，这意味着做一个有能力正确或错误地行为的人，在这个意义上，行为的道德上即相应是好的或坏的”〔32〕。人工智能无论在意志自由抑或道德判断上都无能为力。实践中，人工智能产品的责任承担者往往是在使用者、制造者与监督者之间平衡，但具体如何平衡尚未得出较为权威的规则。人工智能司法

〔26〕 参见郑曦：《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。

〔27〕 翁晓斌、饶淑慧：《人工智能司法决策的可解释性及其路径研究》，载《学习论坛》2022 年第 5 期，第 132 页。

〔28〕 参见〔美〕瑞·L. 马肖：《行政国的正当程序》，沈岿译，高等教育出版社 2005 年版，序言，第 2－10 页。

〔29〕 曹卫东：《交往理性与权力批判》，上海人民出版社 2016 年版，第 126 页。

〔30〕 陈洪杰：《从技术智慧到交往理性：“智慧法院”的主体哲学反思》，载《上海师范大学学报（哲学社会科学版）》2020 年第 6 期，第 90 页。

〔31〕 前引〔9〕，刘艳红文，第 85 页。

〔32〕〔澳〕皮特·凯恩：《法律与道德中的责任》，罗李华译，商务印书馆 2008 年版，第 88 页。

决策责任的模糊问题存在以下两个层面：一是人工智能司法决策错误产生的原因存在多种可能，可能是人工智能代码的错误，也可能是法官操作的失误，还有可能是人工智能技术固有的缺陷；二是人工智能司法决策责任承担的模糊性，由于错误产生背后的原因各异，责任的归属也并非清晰，如人工智能代码的错误或法官操作的失误，这类责任者相对较为明确，但因人工智能技术固有的缺陷所引发的责任，其责任承担者就并非清晰，它有可能是研发者，也有可能是法官，还有可能是国家等。责任者的模糊可能导致司法决策本身的随意与偏见，加大司法者与当事人之间的法律鸿沟，缺乏公平责任的司法体制设计也会使得新技术的应用陷入可解释性“整体危机”。

第六，人工智能无法进行价值衡量，削弱司法正义的基础。人工智能在司法正义判断上的困境也使得算法的可理解性和可接受性大为降低。一者，在法律的适用过程中其无法真正实现“同案同判”，同案本身就是基于两种案件结果价值相似性的判断，而人工智能恰恰无法基于司法正义价值对案情进行整体判断和关照。例如，“于欢案”中涉及对正当防卫制度的重新理解，仅仅将此案编入故意杀人罪一类案件将发生严重的司法不公。换言之，人工智能“无法避免建立‘错误的相关性’，即两个案件尽管具有事实特征上的相似性，但这种相似性却不具有法律意义或不应与法律后果发生关联，而机器学习算法却将其当作了‘链接’法律后果的前提”〔33〕。二者，某些法律价值不存在位阶排序无法进行计算与权衡，这导致算法无能为力。例如，经典的“电车难题”实际拷问的即是生命价值的无可计量性。事实上，在诸多法治之善之间都需要法官依照现实世界的逻辑观和心中世界的正义观作出权益平衡与实质判断，基于相关性的概率计算无法理解大脑神经元“黑箱”，也导致其无法被人脑理解。三者，人工智能有限的数据库和无限的司法要素事物发展之间形成张力。人工智能司法决策是基于已被标记和筛选的过往数据，作旧的数据使得人工智能对新事物的感知能力和关联能力不足，无法应对层出不穷的司法新事物和新形势。例如，部分金融刑事案件的入罪标准需要综合东中西部经济发展状况作出动态平衡，而非采取“一刀切”的算法标准。

三、人工智能司法可解释性的正当逻辑

“我们不可能从对那个时代的详细研究的结果中获知世界大事的意义，即使是这个结果极其完善；相反，我们必须能够创造出意义本身。”〔34〕逻辑分析是客观认识事物意义的前提。人工智能司法可解释性困境的存在说明构建人工智能的可解释性规则存在现实必要性。但这并不等于说构建人工智能司法可解释性规则具有可行性。这首先需要揭示学界关于人工智能司法不可解释性立场的伪科学性，进而构建人工智能司法可解释性规则的逻辑基础。

（一）对人工智能司法不可解释性立场的批判

人工智能实际是计算机科学下模拟“类人智能”所形成的一套计算规则，它的核心技术无疑

〔33〕 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 77 页。

〔34〕 前引〔12〕，马克斯·韦伯书，第 9 页。

是算法，算法被定义为“解决某一特定问题而采取一种有限、确定、有效并适合用计算机程序来实现的解决问题的方法，是计算机科学的基础”〔35〕。学界普遍有一种惯常思维，认为可解释性的最大障碍是算法的“黑箱”存在，其导致人工智能司法决策本身具有“黑洞”空间而无法让人理解。但事实真的如此吗？从逻辑上分析可知，这种立场具有伪科学性。

首先，区分事实和价值、实现客观与主观的有限分离是近现代哲学科学的逻辑起点。在法律适用过程中所形成的算法，虽然经过了人的目的性加工，但本质还是一种运算规则或代码符号，是一种客观产物。就此而言，算法并非像人的主观意志一样，是一种不可重复、复制与展示的“黑箱”，而是一种可探知、重复的客观物质。物质主要有四类特征，即遵守能量守恒定律、具有可探测性、时空有限、遵循因果规律，正是由于人工智能不具有意识这类非物质特征才使得其不能成为人性主体。〔36〕作为一项可重复的运算规则，算法具有高度的形式化特征。算法通常只有三种运算规则，即“是”“非”“或”，其根据相应的技术代码植入，作出相应的精确性指令。就此而言，打破算法“黑箱”或还原算法的自主运算过程在逻辑上是可能的，只不过是技术或成本问题。例如，在司法决策过程中，虽然根据正当防卫这一指令，相应的算法会进行海量数据筛选与深度学习辨识，但得出的结果与基础要素之间总是存在一些线索或访问痕迹，如一些关键性的基础数据要素，包括“侵害”“反击”“急迫”“平衡”等，明确两者之间的关联性需要相当高的支撑技术，但不代表不可能。而这即是物质客观性与实在性的体现，也是可解释性的客观性基础。

其次，按照前述常规思维路径，算法存在“黑箱”等因素，导致或加剧算法的歧视，因此，这种算法歧视是人工智能固有的缺陷，致使人工智能决策的可理解性和可接受性锐减。但实际上，算法决策的“歧视”或“不公”归根结底还是现实社会中制度、规则不公的真实映射。一个很典型的案例即是，如果司法层面习惯报道黑人犯罪事件或选择性公开黑人犯罪数据，那么经过大数据的推演所得出的结论自然是，黑人是犯罪的高概率群体，黑人无疑被贴上犯罪标签。显然这种歧视并非算法“黑箱”所导致的结果，而是社会歧视本身的映射。与此同时，另一种情况是，算法本身是一种中立的统计或分配规则，但人们的固有情愫往往认为一种算法规则比另一种更公平。例如，在美国教育平权案中曾存在两种招生录取算法，个体主义的直接加分规则和特定群体的倾斜加分政策。初看之，对特定群体的倾斜加分政策会比个体主义的直接加分规则更易引发“反向歧视”，但事实不然，从算法结果角度而言，无论是个体主义进路（加分政策），还是群体主义进路（配额政策），都旨在提高某些族裔（特别是黑人）的录取率，都不过是一种“纠偏行动”（affirmative action）。〔37〕就此而言，算法本身并非价值中立或非中立，真正的问题在于如何选取适当的“基于社会效果的法律解释模式与方法”。

最后，退一步而言，即使存在算法“黑箱”，也是因人的社会活动及相应规则所致，而非物

〔35〕〔美〕塞奇威克、韦恩：《算法》（第4版），谢路云译，人民邮电出版社2012年版，第1页。

〔36〕参见程承坪：《人工智能：工具或主体？——兼论人工智能奇点》，载《上海师范大学学报（哲学社会科学版）》2021年第6期。

〔37〕参见丁晓东：《算法与歧视——从美国教育平权案看算法伦理与法律解释》，载《中外法学》2017年第6期。

质（算法）的不可知性。例如，在客体层面，基于司法领域的国家秘密、商业秘密、个人隐私以及知识产权的保护之需要，需要对算法决策系统的数据及规则进行保密，防止黑客攻击、商业剽窃以及隐私泄露的风险，这就导致数据的秘密性；在主体层面，人工智能司法运算场域内的主体知识结构多元，虽然计算机专家熟知算法代码，但法官及诉讼当事人未必理解晦涩难懂的计算机语言，司法算法模型中的许多代码与标量并非像“年龄”和“性别”一样清晰，它往往具有更为抽象的行为序列或模拟符号特征，某些代码与标量也难以用人类语言或可视化图表予以标记。〔38〕在此情况下，一定程度的算法“黑箱”是因这个时代发展过程中主观或客观性制度而延伸的一种“遗产”。真正需要解决的问题是，提高人工智能司法的问责程度和及时提供救济的能力，从而保障当事人的诉讼权利，实现司法公平。利用法律解释和裁判说理等程序方式提高人工智能司法决策的透明性与可理解性，只是直接目的而已。

（二）对人工智能司法可解释性逻辑基础的挖掘

1. 人工智能司法可解释性的认知基础

人工智能司法可解释性的认知基础是指法律文本可通过相应的符号转化成为计算机语言，并促成法律解释到法律解析的飞跃。法律语言转化为计算机语言的核心问题是，法律概念的形式化。而恰巧法律概念具备形式化的条件，这是可解释性的认知基础。

一方面，法律概念形式化的基础在于法律的形式理性，这使得法律规则和算法规则逻辑趋同。在基本结构上，算法主要是一种数字逻辑规则，0 或 1 是其运算的基本参数。而法律为实现规则的精确性与稳定性往往借用数学、经济学、逻辑学等运算规则或符号予以推理论证与表达，如司法机关法律文书效力的公法经济分析，必然涉及数学函数的表达。因此在某种意义上，两种规则实际是置于相同逻辑范式下应用于不同领域的语言体系而已。在运行模式上，算法的认知逻辑无非是基于数据的偏好、节点的分析与预测；在类型上，符号主义认识模式侧重符号逻辑的推理，联结主义认知模式侧重数据单位之间的节点关联性，行为主义认知模式则更加强调强化学习的重要性。与之相比，法律规则在运行模式上也是对主体行为进行提前预设与编入，法律适用的过程大多是一个三段式的自动化应用过程，符合算法符号主义认知模式的基本逻辑。因此，基于本体推理方法的自动化适用是两者共同的运行逻辑。在结果面向上，人工智能在司法裁判领域的运用具有正当性，能够从海量的司法裁判文书中提炼出案件事实、类型、引用的法律法规及判决结论等关键信息，而大部分裁判要素信息都具有类同性与可重复性，通过算法决策自动化能够提高裁判的效率和准确性。就此而言，法学和计算机学都在致力于解决秩序的稳定性问题，计算机法学应运而生。

另一方面，法律概念形式化的表现在于本体要素的提出，这可以实现法律概念知识和计算机算法符号之间的互通。本体要素是指在给定领域内概念的要素化与具象化，以及各个要素对象之间形式化的、明确化的一般性规范结构。通过抽取法律现实生活中概念的本体要素，并以计算机代码标记之，就能对相应的法律规则进行“运算”，得出法律适用的结论。在此，本体要素提供

〔38〕 See Lilian Edwards，Michael & Veale，Slave to the Algorithm? Why a “Right to an Explanation” is Probably Not the Remedy You are Looking for，16 *Duke Law and Technology Review* 18 (2017).

的正是计算机算法运行的“知识概念词汇表”〔39〕。按照本土要素的一种常规提炼方法，首先，需要对一个法律规则进行文本范围分类，确定为哪种法律领域或部门法系统下的法律规范信息；其次，需要对法律规则的规范类型进行认定，如是授权性规范、义务性规范抑或禁止性规范等；再次，需要对法律规范的逻辑结构进行提炼，通常一个完整的法律规则包括假定条件、行为模式和法律后果，有些会省略法律后果或假定条件，但行为模式必不可少；最后，需要对法律规范中存在的一些附加信息进行归类提炼。例如，非结构化信息管理架构类型系统“卢依马系统”，将司法判决按照在法律论证中的功能差异进行层级类型化，一共列出了 9 个层级，分别是对法律规则的“引用”、锁定“法律规则”、寻找“法律裁定或法律裁判”、挖掘“基于证据的事实发现”与“基于证据的中间推理”、提炼“证据”、明晰“法律政策”、形成“基于政策的推理”、得出“特定案例的过程或程序事实”。〔40〕这些法律句子的层级类型化对法律论证检索非常有用，能够对给定的案例文本进行三段论的注解，即案例适用的法律规则是什么、提炼哪些可以支撑法律规则适用的证据以及这些法律规则的渊源来自哪里，由此能快速实现法律论证的自动化搜索。这也说明，司法裁判领域的人机协作具有较大的潜力。

2. 人工智能司法可解释性的制度基础

从现有的法律规范体系来看，可解释性存在较多的制度与规范基础。首先，以知情权、参与权为核心内容的行政参与及信息公开制度是促进可解释性的公法基础。行政参与制度主要体现在行政正当程序的告知制度、阅览制度、听取意见以及异议制度，它是对抗算法权力“黑箱”及异化的主要手段之一。在司法领域，裁判说理制度可以通过对裁判依据、事实、理由及结论的阐释、说明与公示实现裁判过程的公开，裁判文书上网增强了司法裁判的透明性；同时在司法决策过程中存在的最后陈述、当庭宣判、上诉告知等程序机制有利于保障诉讼当事人的参与和知情权。其次，以信息披露、风险预警及责任分担为核心内容的市场合规与监管机制是促进可解释性的私法基础。市场交易强调意思自治与公平竞争，通过告知、披露、风险提示及事后责任等法定义务的设置实现交易双方信息的对称性。例如，在《民法典》《消费者权益保护法》以及金融、证券、医疗、科技等特殊行业的法律法规，都规定了在涉及个人权益、公共利益及市场秩序保护的领域，相应的技术人员、工作人员承担告知、风险警示、信息披露等法定义务。人工智能司法决策系统作为一项市场交易产品也必须遵守市场交易规则，以保护消费者（诉讼当事人）的健康权、知情权及相关权利。最后，以数据权和算法解释权为核心内容的权利保障制度是促进可解释性的混合基础。一方面，数据权是促进可解释性的间接基础。数据权是一项混合型权利，权利的核心在于如何保护数据中包裹着的个人利益信息并协调数据利益冲突形态，“在数据企业对数据要素化利用的普遍期待之外，用户和数据企业同业竞争者对于数据的利益期待之有无，取决于商业模式中所利用的数据是否承载个人信息以及是否处于公开状态”〔41〕，基于“知情同意规则”的

〔39〕〔美〕凯文 D. 阿什利：《人工智能与法律解析——数字时代法律实践的新工具》，邱昭继译，商务印书馆 2020 年版，第 211－212 页。

〔40〕 参见邱昭继：《人工智能、法律解析与未来法律实践》，载《政法论丛》2022 年第 4 期。

〔41〕 沈健州：《数据财产的权利架构与规则展开》，载《中国法学》2022 年第 4 期，第 100 页。

数据利用已成为商业数据产品开发的首要原则，用户对数据的异议权、更正权及删除权可大幅度促进算法的可解释性与透明性。另一方面，算法解释权是促进可解释性的直接基础。一般认为，直接创设算法解释权的法律依据是 2018 年欧盟实施的《通用数据保护条例》（GDPR）第 7 部分“数据主体的权利”下面的“自动化个人决策相关权利”一节第 4 条，其明确指出：“对该条款所涉及的任何处理，都应当采取适当的保障措施，包括向数据主体提供具体信息以及要求人为干预、表达其观点、要求对此类评估后作出的决策进行解释以及质疑此类决策的权利。”国内学者基于此提出了构建本土算法解释权的基本思路及构想，〔42〕通过证成算法的权利属性，增强国家机关、市场责任主体等的信息公开及阐释义务，打开算法“黑箱”歧视的救济通道，从而提高人工智能的透明性。

四、人工智能司法可解释性困境的纾解之道

传统观点认为：“司法的本质是理性，法律推理是一种理性过程，裁决者不能有利益、感情牵涉，中立是最基本的要求。”〔43〕显然，这样的一种观点有夸大司法客观性之嫌疑。正如任何一种法治都有其政治基础一样，任何一套司法体系都是当代社会制度综合的结果。司法作为一种法律运用，既是一种实践理性（practical reason），同样也是人类精神的高级“创造状态”，“是一种不可模式化的实践巧智慧”〔44〕。在理性主义和非理性主义立场之间保持张力平衡正是司法的艺术。人工智能司法在某种意义上有利于保持司法价值判断之客观化，但这又与法秩序规则之主观化存在一定的抵牾。因而，人工智能司法既有必要、也应始终处于客观性与主观性的平衡和交融之中，客观性主要是从制度与机制着手，主观性主要是从人的认识层面着手。解决人工智能司法的可解释性悖论也需要从客观性与主观性平衡和交融的角度来思考。具体包括以下对策：

（一）构建司法信息公开共享制度，提高有用数据的甄别与利用效率

信息公开是人工智能司法可解释性的制度基础。传统的信息公开，一是主要停留在政府管理层面，二是主要强调“知情权”而未突出信息的共享与使用。公共数据共享是一种以数据利用和公平赋权为核心价值目标的公共服务，它突破了政府信息公开在主体、范围和结果上的限制，将其提升至为社会公众创造共同财富的高度。它是一种以“权力—权利”作对位的双向互动法律结构，使国家拥有一定的数据配置权并承担合理利用义务，使公民拥有获取公共数据资源的权利并承担不予滥用的义务。“在这种结构中，国家不是简单地对数据资源进行控制、支配和管理，也不是放任私人自由攫取和使用数据资源，而是建构并维护一种公平合理的数据利用秩序，促进公

〔42〕 参见张恩典：《大数据时代的算法解释权：背景、逻辑与构造》，载《法学论坛》2019 年第 4 期；姜野、李拥军：《破解算法黑箱：算法解释权的功能证成与适用路径——以社会信用体系建设为场景》，载《福建师范大学学报（哲学社会科学版）》2019 年第 4 期；解正山：《算法决策规制——以算法“解释权”为中心》，载《现代法学》2020 年第 1 期；张欣：《算法解释权与算法治理路径研究》，载《中外法学》2019 年第 6 期等。

〔43〕 陈端洪：《司法与民主：中国司法民主化及其批判》，载《中外法学》1998 年第 4 期，第 39 页。

〔44〕 雷磊：《类比法律论证——以德国学说为出发点》，中国政法大学出版社 2011 年版，第 3－4 页。

共数据利用的开放性、公平性、效益性。”〔45〕司法数据共享可转变单向度的司法数据公开现状，将司法数据作为一种有效益的权利资源，促使司法参与者以更加积极的权利主体身份参与司法运行的全过程。同时，权力享有者也应努力提高自己维护公平合理的司法数据共享的意识，仅仅停留在裁判说理制度、裁判文书公开制度等是不够的。一方面应为人工智能司法决策提供更为真实、全面的基础样本，另一方面应为诉讼当事人及民众提供更为准确的司法判决指引。事实上，在既往的刑事案件领域，涉黑、贪污等部分案件与国家层面的刑事政策密切相关，由于存在信息的不对称，部分案件的司法决策过程存在较大的不透明性，这也影响了人工智能在此领域的运用和发展。在新媒体赋权下，部分案件虽通过司法场域的外部系统予以披露，但也会形成基于个别律师单方面发表过激言论的负面情形。倘若司法等职能部门能够采取“充分和具有选择性的开放式许可”〔46〕模式，将传统的司法信息公开转化为公平的数据利用，可以更好地预见人工智能在此领域应用的潜在风险，提高司法决策的效率与可接受性。在具体实施层面，司法信息公开共享制度是前提，制度的最终意义在于实现对有用性数据的甄别、筛选及应用，使之应用于人工智能司法。目前，最高人民法院法院网、中国司法裁判文书网、北大法宝网以及地方法院裁判网提供了不少裁判文书信息，但有效数据仍存在不足。需要联合司法部门、行政部门建立合作机制，推动相应的有用数据的甄别与利用，为人工智能司法提档升级铺路。

（二）从软硬法结合视角建构司法系统的运行标准与制度规则

引发人工智能司法决策不可解释性与多重困境的一个重要原因是，人工智能重塑和颠覆了传统的法律关系结构，引发法律变革，但相应的制度规则及运行标准并未完全建立，如人工智能司法决策模型黑洞问题、人工智能司法决策程序漏洞问题、人工智能司法决策责任模糊问题等，即是由人工智能司法系统的决策标准、程序标准及责任标准与规则未能完全明确而导致。然而，从一个体系层面而言，完善人工智能司法系统的运行标准与制度规则并非仅仅依靠单一的“国家法”手段，还应将“社会法”“行业法”等软法统筹进来，实现硬法和软法的融合、公法和私法的合力。硬法层面主要是设定禁止性规范和义务性规范，这是对算法决策危险的及时禁止和基本权利的救济，主要活跃在公法领域。例如，欧盟《通用数据保护条例》（GDPR）第 22 条第 1 款就对完全自动化决策进行了“一般性禁止”，即“个人有权不受完全依据自动化处理作出的且对其产生法律或类似重大影响的决策的约束”，这也被称为一般性“反对权”，其主要目的是实现“算法控制者不得通过编造的人为干预而规避对自动化决策的一般性禁止，任何名义上或象征性的人工干预均不对自动化决策构成实质性影响”〔47〕。与之相比，美国并未进行统一的立法禁止性规制，而是突出行业规则和依靠法院的算法解释规则。例如，在美国量刑领域的人工智能算法的主要问题是算法解释请求权的法律基础，从现有的依据来看，宪法领域的正当程序权利可作为宽

〔45〕 王锡锌、黄智杰：《公平利用权：公共数据开放制度建构的权利基础》，载《华东政法大学学报》2022 年第 2 期，第 63 页。

〔46〕〔美〕瑞恩·卡洛、迈克尔·弗鲁姆金、〔加〕伊恩·克尔编：《人工智能与法律的对话》，陈吉栋、董惠敏、杭颖颖译，上海人民出版社 2018 年版，第 171 页。

〔47〕 A29WP，A29 WP，Guidelines on Automated Individual Decision—making and Profiling for the Purposes of Regulation，2016/679，17/EN. WP 251rev. 01（Feb. 6，2018），转引自前引〔42〕，解正山文，第 183 页。

泛的权利基础，不仅包括传统的知情权，还包括正当性的解释权，后者是一种真正的“解释权”，即被告可以毫无限制地获取源代码以及算法结果所依赖的逻辑的权利。〔48〕通过已有的法律法规体系及权利请求规制与救济通道，实现权利对权力的规制。软法方面主要是设定行业安全标准和构筑自我规制体系，这是对算法决策风险的及时预防和侵权的救济，主要活跃在私法领域。从人工智能司法解释义务的场景化标准类型来看，它属于公共事业领域的一种衡量标准，相比于一般商业领域而言，在合理性和透明性方面具有更高的要求，在系统上线投入市场使用前，需要经过较为严格的安全性测试，包括安全性、准确性、稳定性、可靠性、保密性等基本要求，同时还必须形成体系化的设计标准、性能标准、运行标准、监管标准、救济标准等。除了行业的自律标准外，算法侵权救济必不可少，即人工智能司法本身出现歧视等大规模侵权事件时，可考虑借鉴欧盟《通用数据保护条例》（GDPR）的代表机构追偿模式，实现群体司法案件的同类化、集约化处理。

（三）从全过程视角强化主体之间的协同治理

人工智能司法决策的治理是涉及多元主体、多个领域的复杂互动与联合规制的议题。因此，不仅需要形成明确的人工智能司法系统的运行标准与制度规则，还需要考虑在司法内部和外部复杂的互动过程中调试出覆盖全过程、全流域的协同治理机制。在聚焦制造主体首要义务、解释模型推进，利用主体审慎义务、解释权的保证，监管主体规制义务、责任公平分担，对象主体反馈义务、合理公平救济的同时，还需要结合内部和外部视角、技术和法律标准、道德和伦理要求、社会和国家互动关系构建一种超越单一权利或权力标准的协同治理机制。其一，从制造主体的角度看，需要强化培育者的信息控制能力与善良动机，从自我规制的角度实现信息与技术的风险控制。问责机制是算法实施透明的重要保障机制，但问责通常具有事后性，通过正面激励和反向问责的双重机制实现培育者的风险自控。其中包括对人员的规制、对算法决策风险管理的目标设定以及完善算法决策风险影响评估制度。例如，欧盟的“AP29 指南”对算法培育主体的治理提出了这样的要求：“企业应当构建有效的内部监督机制，对个人将产生重大影响的算法应当向内部独立的数据保护官提供影响评估的相关信息。同时，企业内部技术团队应当配备专业权威人员对系统的准确性负责，确保其信息可被公众获得，并为救济制度随时启动奠定基础。”〔49〕其二，从利用主体角度看，司法机关需要坚持“技术制衡技术”〔50〕的理念，即提高自身对技术的把握程度，善于将司法规则和理念“代码化”或者“技术化”，将抽象的规制原则或理念转换为实际场景运用规则，防止出现被机器人操控的未知或盲目自信的利用过失。例如，司法机关对人工智能模型的可解释性需要进行审查与评估，可建立由法学专家牵头，吸纳计算机、人工智能、哲学、心理学等多学科交叉领域的专家组成的统一委员会，事前对可解释性的内在透明性进行审查与评估，事后对可解释性的结果保真性与一致性进行认定与审查。其三，从监管主体角度看，主要利用的是一种回应型规制，即将“威慑式规制策略”和“遵从式规制策略”结合起来的混合型规制模式，〔51〕

〔48〕 参见前引〔42〕，解正山文。

〔49〕 张欣：《算法解释权与算法治理路径研究》，载《中外法学》2019 年第 6 期，第 1443 页。

〔50〕 张涛：《探寻个人信息保护的风险控制路径之维》，载《法学》2022 年第 6 期，第 68 页。

〔51〕 参见〔英〕罗伯特·鲍德温、马丁·凯夫、马丁·洛奇主编：《牛津规制手册》，宋华琳、李鸽、安永康、卢超译，上海三联书店 2017 年版，第 134－135 页。

既设置严格禁止、处罚等规则，又强调合作、教育、说服、指导等柔性治理手段，以实现人工智能司法行为的事前、事中与事后规制的全覆盖。其四，从对象主体角度看，要提升自己对人工智能风险的识别，加强沟通，及时作出反馈和救济请求。对象主体主要是指诉讼当事人，也包括律师和其他诉讼参加人，事实上，在线上立案、庭审、结案、电子化文书送达等智慧法院模式建设以来，部分诉讼当事人未能完全适应与转化过来，导致其在智慧法院模式中的被动，甚至不知所云，这自然也会加深对人工智能司法的误解及风险的不可控，因而建立常态化的沟通和反馈机制必不可少。

（四）通过指导性案例和司法解释赋权法官司法解释空间，提高法律解释技术

法律结构的不明确是导致人工智能司法技术难以有效运用的认知障碍之一。虽然在理性实践主义的引领下，法律结构一直被形式化与教义化，但作为法理学上“恼人不休”的话题之一，法律语言的“空缺”和法律结构的开放性，导致司法判例总是存在不确定性。如哈特所言：“英国的判决先例‘理论’对于使用判例实务的理论描述，在某些点上仍旧具有高度的争议：的确，即使在理论当中，‘判决理由’（ratio decidendi）、‘案件事实’（material facts）、‘（法律）解释’这些关键词，也含有不确定的阴影地带。”〔52〕由于不确定性概念和开放性结构的存在，按照对立法冲击程度的排序，法官对法律的适用存在法律解释、法律续造和填补立法几种可能的情形，越是靠后要求法官的能动性越大，因其对现有的法秩序冲击也愈大，受到法体系的限制也越高。我国不是判例法国家，原则上法官适用法律的模式被限定为法律解释和狭义的法律续造，即不可突破法律体系内的限制，主要是对模糊法律语言的一种具体化与情景化解释，以能够让一般民众所理解。实现人工智能司法的可理解性，并非消除法律语言的模糊性和法律结构的不确定性，而是要找到计算机语言和法律语言的衔接点。事实上在形式理性的形塑下，法律规则和算法规则逻辑具有较强的契合性。但这又引发了另一问题，法律不确定性概念和开放性结构如何合理、有效地形式化。法典化是一体解决法律不确定性和开放性的一种手段，但企图一劳永逸的立法又将陷入脱离社会基础的风险。面对这种双向悖论，解决方法一方面是努力增强法律结构的明确性，保持法律规则和算法规则有较高的契合度；另一方面是承认两者之间的裂缝，通过激发法官的创造力、提高法律解释技术，积累更多真实、合理、有效的判决文书，为人工智能学习系统提供实践的“智慧数据”，最大程度弥合裂缝。实践中，通过指导性案例肯定正面的法律解释、赋权法官的司法解释空间，是激发法官的创造力、提高法律解释技术的一种有效途径；在一定的司法解释领域和范围成熟后，还可出台专门的司法解释指导意见，进一步促成法律解释的明确性与透明性。

（五）强化交叉学科人才建设，提高对人工智能司法决策模型的引领

司法算法决策模型设计的合理性、正当性是人工智能司法决策透明性和可理解性的关键要素之一，决定算法决策模式可解释性的核心指标又在于参数的明确程度，参数的明确程度分为模糊型、明显型和外显型三种。例如，在美国的刑事司法实践中 COMPAS 智能量刑模型，不仅采用

〔52〕 前引〔6〕，哈特书，第 199 页。

了标准回归算法，还结合了以实践司法数据筛选为基础的智能学习与分析模型。〔53〕虽然能够打破传统的机械量刑弊端，但对智能算法深度学习的过程和结果可理解性提出了更高的要求。在提高透明性的菜单中有一种溯源机制，即借助区块链技术，将算法过程用分布式账本的形式进行记录，形成过程的不可更改性和结果的可溯源性，解决了运算过程的步骤难以理解的难题。〔54〕目前，我国司法人工智能领域的实践状况是：在参数层面未能充分保障数据的真实有效性，同时缺乏有深度的智能学习与分析司法模型；在解释技术层面也缺少相应的成熟的支撑基础和机制。对于参数基础而言，有学者进行过调研，发现“当下不少所谓的法律科技公司或研究团队严重依赖自己事先假定的知识图谱来提取、印证规范化的裁判模式，其打造的裁判模式可能严重脱离实践模式……稍微复杂的文书识别往往极其困难，因为机器识别在抽取多样、微妙的语言时经常出错，从而影响到大样本材料提取的准确性，最终给出误差很大甚至错误的解读”〔55〕。对于深度学习技术而言，传统法学界习惯于法教义学的推理论证，倾向于定性分析而薄弱于定量分析，学科之间的交融深度有限，很大程度上限制了“面向实践的、统计式的、机器学习介入的研究范式、裁判模型机制”〔56〕的打造。因此，我们加强对人工智能司法决策模型的引领，还需要在参数和深度学习技术层面进行努力，不仅需要更为可靠有效的参数，更需要科学合理的学科交融研究范式、计算机解释及运用技术。在实践层面，一个有效的途径是强化交叉学科人才培养。在目前的学科评估体系中，国内高校，特别是双一流或重点高校可充分利用学位自主审核权以及学科自主评估的机遇，强化交叉学科建设，实现人工智能、法学以及相关学科领域的大融合，为人工智能司法决策模型引领提供人才保障。

（六）发挥法官的自律与能动性，实现司法智能决策的人机协同

在真正的奇点到来之前，人工智能都只是人类意志的一种“延伸”。而真正具有自主意识的人工智能出现时也意味着人类命运将被改写。因此，只有在人类中心主义下人工智能司法才有讨论价值。但是，这也让人工智能司法决策嵌入了司法正义判断困境这一基础性缺陷。由于这一缺陷的根本性，它事实上是难以根除的。司法正义的实现是一个法官运用经验、利益衡量、价值判断、法律解释、制度实施的“司法—社会”互动的过程，“法官在‘司法—社会’的互动中，主要以语言为载体，处理复杂社会关系中的纠纷，司法裁判的形成与其说是是非曲直的判断结果，毋宁说是一场规范与实践之间互动商谈的对话结果，并以此为人们确立了未来的行动标准和行为方向”〔57〕。人工智能的技术即是对人脑海量信息运算不足的弥补，但人工智能必须接受人类规则的规制。在此，经常提及的是人工智能的伦理规制，例如，2017 年“阿西洛马会议”提出的 23

〔53〕 See Megan T. Stevenson & Jennifer L. Doleac, Algorithmic Risk Assessment in the Hands of Humans, available at http://humcap.uchicago.edu/RePEc/hka/wpaper/Stevenson_Doleac_algorithmic-risk-assessment-humans.pdf, last visited on Nov. 25, 2022。

〔54〕 See Ilaria Tiddi, Freddy Lecue, Pascal Hitzler et al., *Knowledge Graphs for Explainable AI—Foundations, Applications and Challenges. Studies on the Semantic Web*, IOS Press, 2020, pp. 243 - 261.

〔55〕 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，第 11 页。

〔56〕 前引〔55〕，左卫民文，第 12 页。

〔57〕 胡铭、宋灵珊：《“人工＋智能”：司法智能化改革的基本逻辑》，载《浙江学刊》2021 年第 2 期，第 22 页。

条人工智能原则。事实上，法律的实施也需要伦理规制，两者都需要人性之善的引领，“有理智的欲望又需要善的价值引导，能够给人们带来好处或益处，而什么是好处或益处则需要有所共识”〔58〕。未来人工智能在人工智能司法决策中完全有可能承担主要作用或主要任务，但司法智能决策的人机协同不在于工作承担量之大小，而在于价值的引领和方向的引导，即遵循国际、国内同行的法律行业标准，坚守司法正义，回应人民诉求。实践中重在对法官进行相应的培训与指导，解决的办法是，对法官进行培训与规训，使其在自律的基础上对人工智能的价值目标不断进行矫正与引领。

五、结语与展望

大数据时代，“人工智能＋”的模式已然成为推动生产力发展的重要手段，人们的生活、工作、环境场域乃至思维方式都在发生大革命。人工智能司法有利于推动法律适用走向一个新的发展阶段、获得新的认知价值。但作为一项计算机认知技术和推理规则，它在面对法律的模糊性、开放性和价值性时不免遇到形式化规则和实质化正义之间的张力。人工智能司法可解释性悖论是这种张力外化的一种表现，消解人工智能司法的可解释性悖论也成为时代发展不可回避的课题之一。人工智能司法可解释性问题主要涉及基础数据、目标任务、算法模型以及人的认知这四类关键要素。但从理论层面而言，人工智能司法可解释性悖论是一个伪命题，这种悖论可以从认知基础和制度基础两个方面进行消解。基于认知基础和制度基础的可解释性基础，也衍生出客观主义和主观主义两种消解悖论的路径，具体包括：构建司法信息公开共享制度，提高有用数据的甄别与利用效率；从软硬法结合视角建构司法系统的运行标准与制度规则；从全过程视角强化主体之间的协同治理；通过指导性案例和司法解释赋权法官的司法解释空间，提高法律解释技术；强化交叉学科人才建设，提高对人工智能司法决策模型的引领；发挥法官的自律与能动性，实现司法智能决策的人机协同。

当然，人工智能司法在运用的过程中要考虑全国各地的实际情况。我国是一个幅员辽阔、区域差异化较大的社会主义大国。大国治理既要注重法治统一，又要注重地区差异。人工智能司法技术在各地区运行过程中，必然会存在技术差异、场景差异以及案件类型的差异。例如，在技术层面，东部地区更为先进与普及，而西部地区相对落后和短缺；在场景层面，在司法数据收集、司法辅助审判和司法决策中算法的精确性存在差异；在类型层面，刑事案件、行政案件以及民事案件对证据的关联程度以及因果链条的要求逐渐降低，人工智能决策可解释性与透明性的要求也有所差异，同时各地方在应用人工智能司法的重心存在差异，如云南地区比内部省份对毒品犯罪的司法智能审判更高。基于这些差异，人工智能应当根据具体技术水平、场景要求、案件类型等因素对司法进行差异化介入，而非打造普适性人工智能司法决策系统，在全国范围内无差别性地推广。

〔58〕〔美〕阿拉斯代尔·麦金泰尔：《现代性冲突中的伦理学：论欲望、实践推理和叙事》，李茂森译，中国人民大学出版社 2021 年版，第 11 页。

人工智能司法治理是一项长期工程，需要以回应时代需求的态度加强人工智能司法决策解释机制、运行程序、问责制度、影响性评估、监管组织架构等核心议题的研究，同时深化司法改革步伐，注重法律价值和技术理性的平衡，以实现占领未来人工智能司法高地的战略目标。

Abstract: Strengthening the research on the development and risk of artificial intelligence justice is a topic of the times, in which the interpretability dilemma of artificial intelligence justice is particularly critical. AI judicial interpretability refers to the comprehensibility and transparency of judicial decisions or behaviors, involving four key elements: basic data, target tasks, algorithm models and human cognition. Unexplainable dilemma is mainly caused by factors such as data failure, algorithm black box, limitations of intelligent technology, decision-making procedures and lack of value. However, the unexplained paradox of AI justice is actually a false proposition, which has two aspects of cognitive and institutional basis. The specific strategies to relieve the dilemma include building a judicial information public sharing system and improving the screening and utilization efficiency of useful data, constructing the operation standards and system rules of the judicial systems from the perspective of the combination of soft and hard laws, strengthening the collaborative governance between the subjects from the perspective of the whole process, empowering judges with judicial interpretation space and improving legal interpretation technology by guiding cases and judicial interpretation, strengthening the construction of interdisciplinary talents and improving the guidance of AI judicial decision-making model, giving judges scope for their self-discipline and initiative, and realizing the human-computer coordination of judicial intelligent decision-making. In the future, it is not only necessary to grasp the balance between judicial value and technical rationality, but also need to consider the differential intervention of AI in justice, so as to promote the realization of AI judicial strategic objectives.

Key Words: artificial intelligence, justice, algorithm, explainability, collaborative governance

(责任编辑：赵　真　赵建蕊)