

论数字司法价值对齐的实现路径

彭中礼*

内容提要 数字司法中算法决策与司法价值的偏离容易引发伦理与法律风险。在数字司法场域中推进价值对齐,应当坚守司法本质以定其向,防控技术风险以固其基,筑牢司法信任以彰其效,融合司法文明以臻其境。数字司法价值对齐要求算法的价值追求与法律精神、社会伦理、人文关怀形成动态适配,其具体要求体现为维护一致性、确保可控性、保持鲁棒性以及避免恣意性。推进数字司法价值对齐,需凝聚全链条、全过程的协同合力,在源头校准层、运行管控层、迭代优化层与安全保障层系统构建层次化推进的过程机制。理论分析表明,数字司法中的价值对齐通过将人类多元价值转化为技术可理解、可执行的逻辑,让人机协同从效率适配的工具层面跃升为价值共生的生态层面,最终形成技术进步与人类发展深度契合的良性循环。

关键词 数字司法 价值对齐 人工智能 司法价值 算法正义

DOI:10.14111/j.cnki.zgfx.2026.03.005

目次

- 一、引言
- 二、数字司法价值对齐的逻辑机理
- 三、数字司法价值对齐的规范要求
- 四、数字司法价值对齐的过程机制
- 五、结语

* 中南大学法学院教授。本文系 2025 年度湖南省社科基金重点项目“风险社会视域下国家预防型义务的配置原理与制度因应”(项目批准号:25ZDB010)的阶段性成果。

一、引言

近年来,人工智能等现代数字科技在司法领域的应用,已从单纯的司法辅助工具逐步向深度嵌入司法决策的方向演进。2024年12月,《人民法院第六个五年改革纲要(2024—2028年)》提出“完善数字法院建设体系”,以数字法院建设为代表的数字司法发展进入新阶段。⁽¹⁾在数字司法中,新科技的深度应用打破了传统司法对物理场所、专业中介机构、人类决策者的依赖,突破了既有流程的局限与固有目标的束缚,⁽²⁾已然掀起司法转型的全新浪潮。

但是,数字科技在司法领域的深度应用,也带来了一系列亟待回应的伦理与法律难题。⁽³⁾算法的纵深渗透与技术创新的加速迭代,进一步加剧了算法逻辑与法律体系之间的结构性紧张关系。其中尤为突出的是,算法歧视、黑箱决策等典型问题充分暴露出人工智能在理解法律原则、权衡司法价值等方面存在缺陷,容易导致“合法但不合理”的价值困境。因此,需要在算法设计与应用中积极稳妥推进价值对齐,将法律原则、司法价值嵌入算法逻辑,实现算法运行与司法目标的协同一致。

当前,哲学、伦理学有关价值对齐的研究较为丰富,主要聚焦在三个方面:一是从伦理的角度对价值对齐进行研究;⁽⁴⁾二是对价值对齐的概念、应有内涵以及实施路径进行讨论;⁽⁵⁾三是对人工智能或者大模型是否需要价值对齐进行讨论⁽⁶⁾。法学界也已经开始关注价值对齐:一是对生成式人工智能的价值对齐问题进行法理反思⁽⁷⁾,探讨法律原则作为价值对齐的标准⁽⁸⁾;二是对价值对齐的法治进路进行宏观讨论⁽⁹⁾,试图构建法治化机制⁽¹⁰⁾;三是基于具体的部门法探讨价值对齐的实现⁽¹¹⁾,关注自动驾驶过程中的个性化价值对

(1) 学界与实务界对数字司法主要存在两种认知界定:第一种为广义理解,即数字技术赋能公安机关、人民检察院、人民法院全司法链条履职,由此形成一体化新型司法运行模式;第二种为狭义理解,是指数字技术在人民法院审判工作中的智能化应用模式。本文主要从狭义理解的角度展开分析。

(2) See Ethan Katsh & Orna Rabinovich-Einy, *Digital Justice: Introduction*, International Journal of Online Dispute Resolution, Vol.3:102, p.123 (2016).

(3) 参见郑智航:《人工智能算法的伦理危机与法律规制》,载《法律科学》2021年第1期,第14-26页;孙保学:《人工智能算法伦理及其风险》,载《哲学动态》2019年第10期,第93-99页。

(4) 参见张逸霄、蓝江:《价值对齐的现实问题与中国方案》,载《思想理论教育》2025年第5期,第29-36页;沈湘平:《价值对齐与人类价值共识及其生存理性》,载《自然辩证法研究》2024年第12期,第3-11页。

(5) 参见闫宏秀:《基于信任视角的价值对齐探究》,载《浙江社会科学》2024年第6期,第39-48页。

(6) 参见宋保林:《人工智能大模型价值对齐的伦理建构》,载《伦理学研究》2025年第3期,第94-99页。

(7) 参见范进学、陈阳:《生成式AI价值对齐的法理反思与路径选择》,载《南京社会科学》2024年第10期,第69-81页。

(8) 参见黎江虹、卢希希:《法律原则作为人工智能价值对齐的标准》,载《荆楚法学》2026年第1期,第101-111页。

(9) 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,第75-91页;刘祖兵:《大模型价值对齐的批判、修正与法治因应——基于DeepSeek行为风险的分析》,载《北京社会科学》2025年第10期,第106-116页。

(10) 参见罗有成:《人工智能大模型价值对齐的法治化机制》,载《北方法学》2025年第6期,第50-63页。

(11) 参见唐林垚:《公司法如何促进模型可信与价值对齐》,载《东方法学》2024年第2期,第76-87页。

齐⁽¹²⁾和智慧税务中的价值对齐⁽¹³⁾等问题。然而,现有关于价值对齐的研究尚未充分关注数字司法场景,对数字司法中价值对齐如何实现的问题更是缺少系统性分析,以至于当前的数字司法建设缺乏充分的智识支撑。为此,本文首先剖析数字司法中价值对齐实现的逻辑机理,并揭示其规范要求,继而系统分析实现价值对齐的过程机制,为数字司法有效推进价值对齐提供可能的路径指引。

二、数字司法价值对齐的逻辑机理

数字司法价值对齐,是指在司法数字化、智能化过程中,通过技术嵌入、制度校准、伦理约束,确保算法模型、数据治理与数字司法系统的运行目标、逻辑和结果,与司法核心价值以及人类社会主流价值观保持一致、协同适配的理念与机制。在数字司法中坚持价值对齐,既符合技术赋能的实践逻辑,也契合筑牢数字司法正当性根基的理论逻辑。为此,数字司法要通过价值对齐坚守司法本质以定其向,防控技术风险以固其基,筑牢司法信任以彰其效,融合司法文明以臻其境。

(一) 坚守司法本质

价值对齐及其所凸显的人类价值共识问题,从根本上要求我们深刻把握并始终坚守全人类共同价值。这种共同价值“以共在为前提,以共感为基础,迈向主体间性,运用横向理性,构建公共理性”⁽¹⁴⁾。当然,数字司法价值对齐并非单一维度的价值适配,而是既要锚定人类共同价值,又要紧扣司法价值的本质要求,从而契合数字司法的特定价值导向。

第一,数字司法要通过价值对齐锁定司法价值,坚守司法的本质属性。司法价值如同深嵌司法制度的“基因”,既决定司法的本质属性,更引领司法制度的演进方向。传统司法向数字司法转型,是技术手段的升级,而非价值基础的重构。数字司法作为司法制度的数字化转型形态,需以司法正义为核心价值引领,才能避免技术工具理性凌驾于司法价值理性之上。所以,人们需要将司法价值确立为技术应用的根本准则。数字司法若放弃对司法价值的维护,就会沦为技术的附庸,导致其结果与社会公众的价值预期脱节。

第二,数字司法要通过价值对齐防止价值遮蔽,实现司法的价值引领。司法问题一旦转化为技术路径,就容易形成事实上的技术权力。这样,“算法超越其原本的工具属性,获得了实质调配资源的力量”⁽¹⁵⁾。这种权力形态容易遮蔽司法原有的核心价值内容,造就“算法利维坦”,⁽¹⁶⁾从而可能导致司法价值的偏离。数字司法价值对齐,是要推动司

(12) See Wang Peiran, *Jurisprudential Logic and Institutional Pathways of Personalized Value Alignment in AI*, China Legal Science, Vol.12:109, p.109-132 (2024).

(13) 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,第28-38页。

(14) 前注①,沈湘平文,第3页。

(15) 宋宗宇、杨易川:《伦理价值嵌入算法治理的机制保障》,载《西南民族大学学报(人文社会科学版)》2025年第6期,第50页。

(16) 参见张爱军:《“算法利维坦”的风险及其规制》,载《探索与争鸣》2021年第1期,第95页。

法数字化进程深度契合人民群众对司法的根本期待,因而需始终锚定并坚守人的核心价值本位。

第三,数字司法要通过价值对齐约束技术理性,明晰司法的价值边界。数字司法必须确立清晰的价值边界,这既是由司法的本质属性与数字科技的工具理性之间的内在张力所决定,也是确保数字司法不偏离其根本目标的逻辑必然。尽管技术迭代日新月异,数字社会下的纠纷样态也层出不穷,但司法公正、权利保障等核心价值始终是不可逾越的底线。唯有通过价值对齐锚定技术应用的方向,才能让数字科技始终扮演司法的赋能者而非主导者,确保数字司法在时代变革浪潮中守住司法初心。

(二) 防控技术风险

贝克认为,“风险已经是工业时代的特征”⁽¹⁷⁾。数字司法在运行过程中存在数据安全、隐私泄露等风险。价值对齐在技术维度的核心意义在于,通过技术架构的优化,将抽象的司法价值转化为可量化、可执行的技术规则,从而为防控技术风险带来新的思路。

第一,数字司法要通过价值对齐防控技术异化带来风险。数字技术天然带有工具理性取向,对司法的伦理维度与价值属性缺乏内在敏感。若缺乏制度对核心价值的维护,技术极易依循自身发展逻辑而走向异化。技术异化不仅危害物质文明,亦会侵蚀精神文明,⁽¹⁸⁾其原因在于人类对现代技术难以实现完全的支配与规训。当技术异化传导至司法领域,便容易导致数字司法对司法核心价值的偏离,因此必须通过科学合理的制度设计加以规范与约束。

第二,数字司法要通过价值对齐实现风险防控整体协调。数字司法虽然是人工智能、大数据等技术在司法领域应用的结果,但要建好数字司法,绝非依赖单纯的“技术叠加”,也需要多主体形成整体协同的发展合力。尤其是在数字司法运行体系中,算法设计者、算法监督者、司法办案人员等核心主体,必须锚定司法公正等核心价值实现价值对齐。进而言之,应统筹协调各类司法参与主体,凝聚统一的司法价值共识,再通过价值对齐机制将共识内嵌于数字司法全流程,以统一价值理念引领、规范和指引数字时代的司法实践,筑牢数字司法风险防控的价值根基。

第三,数字司法要通过价值对齐防止风险防控无差别对待。在人工智能、大数据等技术的发展大潮中,不同性质的算法可能带来不同程度的风险。如对其采取同质化防控,易在数字司法中牺牲司法的核心价值。如,算法常被用于案件分流、证据分析、量刑辅助等环节,其中既存在算法歧视这类偏离法律面前人人平等原则的共性风险,又存在因应用环节不同而风险强度与表现也不同的个性风险。这要求价值对齐必须精准化,依托价值对齐识别数字司法各阶段的具体风险,实现风险有效防控与司法价值坚守的有机统一。

(三) 筑牢司法信任

信任是建立社会秩序的主要工具之一。⁽¹⁹⁾在数字司法领域,价值对齐是构建社会信

(17) [德] 乌尔里希·贝克:《风险社会:新的现代性之路》,张文杰、何博闻译,译林出版社2018年版,第6页。

(18) 参见郭冲辰、陈凡:《技术异化的价值观审视》,载《科学技术与辩证法》2002年第1期,第1页。

(19) 参见郑也夫:《信任与社会秩序》,载《学术界》2001年第4期,第30页。

任,进而保障其可接受性的核心路径。数字司法可接受性的实现,并非单纯依赖国家强制力,更根植于社会公众的信任与认同。若数字司法脱离价值对齐的导向,极易使其丧失公众信任,最终侵蚀司法权威根基。

第一,价值对齐要契合社会大众对司法的价值期待,从而构建公众对数字司法本质的信任。公众对司法的信任,首先源于对司法本质属性的确认,即社会大众判断司法工作是否坚守司法公正、程序正当等核心价值,能否切实维护这些价值。通过司法实现公平正义,是社会正义理论体系探索的一个基本命题。⁽²⁰⁾ 数字司法若一味强调技术效率,忽视对司法价值的回应,极易引发公众对技术异化司法的担忧,进而否定其可接受性。价值对齐的核心作用,正是将抽象的司法核心价值,转化为数字司法中可计算、可呈现的具体形态,让公众清晰感知到,数字工具并未改变司法以人为本的本质,从而筑牢社会大众对数字司法的价值信任根基。

第二,价值对齐要契合社会大众对司法技术风险防范的期待,从而构建公众对数字司法安全的信任。技术的迅速发展,确实给人类带来很多便利,但是也存在很多风险。虽然工业社会由其自身系统制造的危险而突变为风险社会,⁽²¹⁾但人民的安全仍至关重要。在数字化背景下,公众对数字科技介入司法领域抱有安全隐忧,隐私侵犯、数据泄露、算法偏见等问题极易引发信任危机和认同障碍。若无法以价值共识化解此类风险,即便数字司法大幅提升效率,也会因安全隐患失去社会认同。价值对齐以安全为重要导向,通过植入价值考量的风险防控机制,将安全要求深度嵌入数字司法的全流程,能有效消解公众焦虑,为数字司法筑牢安全层面的信任根基。

第三,价值对齐要契合社会大众对司法技术价值共识的期待,从而构建公众对数字司法认同的信任。司法的可接受性,既源于对法律规则的严格恪守,又依托于对社会普遍伦理共识的有效契合。在平等参与和协商交流的当代社会中,共识的缺位将使得司法过程缺乏基础的正当性。⁽²²⁾ 数字司法若脱离社会共识,即便严格依法裁判,也可能出现合法不合理的情形,进而引发公众质疑。价值对齐的重要功能正是将动态发展的社会共识融入数字司法的运行逻辑,增强司法的社会认同与可接受性。没有价值对齐,数字司法可能沦为效率至上的“技术玩具”,即便技术再先进,也会因缺乏信任支撑而难以落地。而通过价值对齐构建的社会信任,能让数字司法真正突破技术桎梏,成为公众可接受、可认同的数字时代正义的新载体。

(四) 融合司法文明

司法文明是司法活动及其制度在历史演进中形成的进步形态与文明程度,是法治文明的核心组成部分。数字司法虽是技术与制度融合的产物,但也承载着社会道德共识与人文关怀,为其价值对齐提供了根本性精神指引。这一精神内核能够有效规约算法应用

(20) 参见杨一平:《司法正义论》,法律出版社1999年版,第1页。

(21) 参见[德]乌尔里希·贝克:《世界风险社会》,吴英姿、孙淑敏译,南京大学出版社2004年版,第102页。

(22) 参见孙光宁:《司法共识如何形成?——基于判决的可接受性》,载《山东大学学报(哲学社会科学版)》2010年第1期,第105页。

边界、锚定技术赋能方向,深度契合司法文明发展脉络,有助于构建现代化司法文明体系。

第一,数字司法要以价值对齐彰显司法温度,这是数字司法融入司法文明的核心方向。司法是承载价值判断的实践活动,绝非技术逻辑的简单叠加。如习近平总书记所说,“司法活动具有特殊的性质和规律,司法权是对案件事实和法律的判断权和裁决权”⁽²³⁾。数字技术在提升司法效率、优化流程管理的同时,其“冰冷”的理性特质易使司法过程趋于机械化、程式化,消解司法本应蕴含的人文关怀与温度。因此,数字司法建设必须强化对伦理维度的审视与把控。在算法设计、流程构建、规则制定等全环节,需将公平正义、人权保障、程序正当等核心司法价值作为根本遵循,实现技术逻辑与价值理性的深度融合。

第二,数字司法要以价值对齐维护隐私、保障尊严,这是数字司法融入司法文明的底线要求。在数据驱动的司法模式下,当事人的个人信息、生物特征及生活轨迹等数据被高频采集,若单纯追求技术效率而漠视隐私保护,极易背离司法尊重个体尊严的初衷。为此,数字司法的伦理建构必须通过价值对齐,严守“最小必要原则”,严格限定数据采集范围,禁止司法数据流入非司法场景,从源头防范技术对权利的侵蚀。通过价值对齐的制度化落地,确保数字技术在提升司法效能的同时,始终承载人文关怀,筑牢司法伦理的坚固防线。

第三,数字司法要以价值对齐促进司法文明导向的技术动态迭代,这是数字司法融入司法文明的关键保障。有人可能会产生疑问,数字司法价值对齐究竟是通过技术优化实现“看得见的正义”,还是通过伦理嵌入重构司法与社会的互动、赋予司法文明新的时代内涵?笔者认为,数字化深刻重塑人类生活,必然推动司法文明演进。数字技术与司法文明的融合并非静态适配,而是随技术革新与文明发展持续调适的动态过程。价值对齐作为维系技术与司法文明同频共振的核心纽带,以司法文明的核心要义为根本遵循,将有效打通技术研发与司法实践之间的壁垒,最终助力实现司法文明的迭代升级与提质增效。

三、数字司法价值对齐的规范要求

一般意义上的价值对齐,强调人工智能应当秉持“与人类一致的价值观”⁽²⁴⁾。在此语境下,数字司法以人类共同价值为对标方向,精准契合价值一致性的核心要义。推进数字司法价值对齐,先要明确其规范基准,确立其底层标尺与正当性根基,再进一步细化推进价值对齐的具体要求,为数字司法价值对齐的落地实施提供清晰明确的行动指引。

(一) 数字司法价值对齐的规范基准

在数字司法语境下,价值对齐需契合司法价值尤其是数字司法价值的特殊性。一方面,数字司法的价值导向需深度契合人类社会的普遍共识,与人类社会普遍的道德判断、

(23) 习近平:《加快建设公正高效权威的社会主义司法制度》(2014年1月7日),载《习近平法治文选》(第1卷),中央文献出版社2025年版,第45页。

(24) Martin. Peterson, *The Value Alignment Problem: A Geometric Approach*, Ethics and Information Technology, Vol.21:19, p.19 (2019).

核心伦理认知保持同频共振;另一方面,数字司法的运行全流程更需借助价值对齐机制,充分彰显司法本身的应然价值内核。因此,数字司法价值对齐的规范基准是数字司法中的司法价值。

1. 数字司法价值对齐要对齐司法价值

有学者认为:“司法的元价值主要有秩序、正义、法治与人权。”⁽²⁵⁾而秩序、正义、法治、人权等司法元价值,本质上就是法律价值。司法固然应当遵从这些传统价值,但数字司法所遵从的并非单纯的传统法律价值,而是在司法特定场景中经过改造的价值,可概括称之为司法价值。具体可从以下三个维度进行解释:

第一,数字司法的核心场域是司法实践,价值对齐需适配司法权运行的特殊规律。数字司法是司法权在数字时代的运行形态,是技术进步在法律系统中引发的重大变革,其场域限定于立案、审理和裁判等司法活动。价值对齐的核心目标是解决技术逻辑与司法权运行逻辑的适配问题,而非对整体法律价值的全面映射。

第二,数字司法的核心功能是实现司法公正与效率的统筹推进,价值对齐需直接锚定司法的价值目标。法律价值的功能是为整个法律体系提供价值引领,其目标具有宏观性与间接性;而数字司法的核心功能是通过技术赋能优化司法过程、提升司法质量。数字司法价值对齐将防止技术工具偏离司法的目标,避免技术工具侵蚀司法公正。

第三,司法价值是法律价值在司法场域的具象化与实践化,价值对齐需具备可操作性。法律价值与司法价值是本源与派生、指导与践行的关系。一方面,法律价值是司法价值的本源与核心指引,它为司法价值划定了根本边界与价值基准;另一方面,司法价值是法律价值的实践转化与特殊呈现。司法价值是法律价值在司法场域的派生形态,聚焦司法活动的具体目标与准则。

2. 数字司法价值对齐要凝聚数字司法价值

数字司法价值对齐的司法价值是在数字司法中能够实现的数字司法价值。数字司法价值是司法价值在数字场景中的具象化体现,是新科技应用于司法领域时,所要实现的符合司法本质要求的价值追求。数字司法价值与传统司法价值之间存在承续与分野,进而彰显出自身特有的理论内核与实践品格。

第一,数字司法价值以传统司法价值为底色,叠加数字正义的新内涵。传统司法价值以公平正义、程序正当等为核心基准,该基准构成数字司法的价值底色。但数字司法的价值基准并非对传统司法价值基准的简单复刻与移植,而是深度融入包括数据资源合理分配、数字权利充分配置、算法决策公开透明和代码规制规范有效等在内的数字正义特殊意涵,⁽²⁶⁾形成“传统基准—数字维度”的复合价值结构。

第二,数字司法价值以传统司法价值为根基,深度依赖数字化新载体。传统司法价值的实现,主要依赖“人力+制度”的双重载体。法官的专业素养、职业道德是实现司

(25) 江国华:《中国司法学》,武汉大学出版社2016年版,第247页。

(26) 参见周尚君、罗有成:《数字正义论:理论内涵与实践机制》,载《社会科学》2022年第6期,第166页。

法公正的核心人力保障,诉讼法、证据规则等则是规范司法行为、保障价值实现的制度支撑。而数字司法价值的实现,形成了“技术+数据+制度”的多元复合载体。其中,智能系统成为司法价值实现的重要中介,司法数据则成为驱动司法价值优化的核心要素。

第三,数字司法价值以传统司法价值为遵循,呈现数字司法价值冲突新形态。传统司法价值冲突,多表现为形式正义与实质正义、效率与公平等经典维度的张力,其消解路径主要依靠法官行使司法裁量权、运用法律解释方法予以调适平衡。而数字司法领域的价值冲突既涵盖传统司法固有的冲突样态,更衍生出技术逻辑(效率优先等)与司法价值之间的特殊对立,其消解路径亦需深度嵌入技术规制理念与规则。

3. 数字司法价值对齐中司法价值的双层内涵

数字司法价值对齐的司法价值,既要涵盖传统司法核心价值,也要囊括数字技术赋能司法实践所衍生的独特价值,即数字司法价值。传统司法价值与数字司法价值,共同构成了数字司法发展中必须深度融合的价值共同体,是数字司法价值对齐必须遵循的规范基准。

一方面,传统司法价值相对固定,主要围绕公正、权利保障、程序正当等核心要素展开,这些要素均根植于司法权的权力属性与社会治理功能。其中,公正是司法的首要价值,是司法权运行的根本准则,确保纠纷解决的结果公平合理;权利保障是司法的核心使命,聚焦于维护公民合法权益,彰显司法的人文关怀;程序正当是司法公正的重要支撑,通过规范司法流程,保障诉讼参与人的各项权利。这些司法价值并非主观建构,而是依托司法权的公权力属性,结合司法的纠纷解决、社会治理等职能,在长期的司法实践中逐步沉淀而成,构成司法活动必须恪守、不可逾越的价值底线。

另一方面,数字司法在承续传统司法价值的基础上,添加了技术安全、数据伦理、透明公开等新兴价值维度,这些维度成为数字司法价值体系的重要组成部分。技术安全价值主要体现为保障数字司法系统的稳定性、安全性与可靠性,避免因技术故障、网络攻击等导致司法程序中断、数据泄露或裁判错误。数据伦理价值聚焦于规范司法数据的采集、存储、使用与共享,⁽²⁷⁾避免数据滥用对当事人隐私权等合法权益造成侵害。特别是司法数据包含大量当事人的个人信息、案件敏感信息,若缺乏规范约束,可能出现数据泄露、过度采集等问题,违背司法保障权利的核心要求。透明公开价值是数字技术赋能下司法公开原则的升级与延伸,其核心是依托数字化手段,打破司法流程中的信息壁垒,实现司法活动全环节、全要素的可查询、可监督、可追溯,既坚守传统司法公开的核心要义,又结合数字技术特性形成独特的实践形态。

(二) 数字司法价值对齐的具体要求

价值对齐作为数字司法健康发展的核心逻辑,需以规范基准为出发点,推动数字司法中的算法价值追求与法律精神、社会伦理及人文关怀深度契合、同频共振,其具体要求集中体现为维护一致性、确保可控性、保持鲁棒性、避免恣意性四个方面。

(27) See Simon Vydra et al., *Big Data Ethics: A Life Cycle Perspective*, Erasmus Law Review, Vol.14:24, p.24-44 (2021).

1. 维护一致性

算法深度融入司法运行全过程,催生了以技术赋能司法、数据驱动正义为特征的数字司法新形态。在这一阶段,“技术深度赋能司法体现为最大化司法裁判的可计算空间”⁽²⁸⁾。但数字司法中的算法是天然的“技术派”,可能缺乏对司法价值的自主感知与主动承载,为此,需要通过价值对齐实现算法价值与司法价值的契合。

第一,通过价值对齐实现数字司法追求程序公正与实体公正的统一。关于程序公正和实体公正的关系,有学者认为二者可能有冲突,⁽²⁹⁾但更多学者主张二者是统一的⁽³⁰⁾。程序公正和实体公正的统一在数字司法中要得到充分体现,一方面,算法设计要能够将司法流程环节标准化,保证司法参与人均能够获得相应的程序性权利;另一方面,算法设计也要坚持排除非理性干预,实现规则统一适用,减少人为裁量的随意性,为司法公正筑牢技术屏障。

第二,通过价值对齐实现数字司法追求司法效率与司法公正的统一。司法的效率与公正问题,在市场经济时代尤获关注。在市场经济活动日益频繁的现代社会,如何在确保司法公正的前提下更快提高司法效率,是世界各国司法程序所关注的焦点。⁽³¹⁾在数字司法中,一方面,算法要能够促进资源优化以契合司法繁简分流、快速解纷的现实需求,并帮助司法文书快速生成;另一方面,运用算法提升司法效率不得偏离司法的本质,不能违背司法公正的要求。

第三,通过价值对齐实现数字司法追求司法普遍性与司法个性化的统一。在保障司法普遍性上,数字司法依托价值对齐中统一法律适用的核心要求,通过标准化技术架构消解差异。依托统一的智能办案平台,将法律条文、司法解释、类案裁判规则转化为算法模型,为不同地域、层级的司法主体提供一致的法条检索、类案推送服务。同时,在落实司法个性化上,数字司法以价值对齐精准推进权利保障为导向,⁽³²⁾通过技术精准赋能实现差异化裁量,使二者在公正、权利保障的核心价值引领下达成统一。

2. 确保可控性

数字司法所运用到的大模型,是人类智能技术赋能司法实践的结果。在数字科技深度融入司法领域的进程中,价值实现的可控性已成为防范数字司法技术异化的核心前提,而价值对齐通过将数字司法的技术逻辑与司法价值深度“捆绑”,正成为构建这一可控性框架的关键路径。

第一,通过价值对齐实现数字司法中算法价值取向的可控性。数字司法价值对齐并非对技术应用的绝对管控,而是关于数字司法价值的制度设计与技术校准之间的动态平

(28) 雷磊:《数字司法的理论反思:意义、问题与监管》,载《交大法学》2024年第6期,第86页。

(29) 参见[英]戴维·米勒:《社会正义原则》,应奇译,江苏人民出版社2005年版,第140页。

(30) 参见陈光中:《坚持程序公正与实体公正并重之我见——以刑事司法为视角》,载《国家检察官学院学报》2007年第2期,第3页;樊崇义:《司法要追求司法公正与司法效率的统一》,载《人民法治》2016年第10期,第77页。

(31) 参见前注,樊崇义文,第77页。

(32) 参见高莹:《论智慧法院中的数字正义》,载《江西社会科学》2023年第8期,第164-165页。

衡。确保数字司法的运行逻辑始终围绕司法价值展开,避免因技术自主迭代、效率优先导向引发数字司法中的算法价值偏离,是价值对齐的首要任务。人应当始终是机器的尺度。⁽³³⁾ 数字司法中的人工智能可依托数据与算法辅助司法流程和司法决策,但绝不能突破工具属性这一定位,更不能上升为司法价值观念的最终决定者。

第二,通过价值对齐实现数字司法中算法价值内容的可控性。价值对齐中所要对齐的价值内容,需要由人类来定义。主体价值的内容是主体的创造力,它是主体创造客体价值的能动力量。⁽³⁴⁾ 人作为主体,一切价值均由人生产且契合人的需求,因此数字司法中,价值必须由人类定义和评判。司法价值并非抽象规则,而是植根于社会伦理共识、法律精神与个案情理的有机整体,需法官结合个案裁量、回应当事人诉求、契合社会动态期待,而非让人工智能反向塑造司法的价值内核。

第三,通过价值对齐实现数字司法中算法价值基准的可控性。数字司法价值对齐要求算法运行始终契合人类司法固有的价值准则,这既是司法活动的本质要求,也是数字司法必须坚守的价值前提。司法固有的价值基准,既回答了司法为何存在的根本命题,也为司法运行划定了不可逾越的边界,更是数字科技深度融入司法、防范技术异化的根本遵循。因此,技术在司法中的赋能水平越高、影响力越大,对价值引领的依赖与紧迫性就越强。唯有坚持以公平、正义、人权、程序正当等核心价值引领技术应用,让科技服务司法而非主导司法,才能实现数字赋能与司法价值同频共振,使数字司法成为守护社会正义、保障人民权利的坚实支撑。

3. 保持鲁棒性

鲁棒性是一个跨学科词汇,指的是既保持稳定但又具有一定灵活性的状态。在不确定性和危机出现的情况下,鲁棒性已经成为系统能否生存的关键,⁽³⁵⁾ 因为鲁棒性证明系统有提供值得信赖的服务的能力,能够在特定条件下保持系统/制度稳健运行⁽³⁶⁾。在数字司法中,通过价值对齐实现算法价值的鲁棒性,是让算法在数据变化、场景迭代、社会价值共识微调的复杂条件下,始终保持对司法价值基准的稳定遵循。

第一,通过数字司法价值对齐保持司法价值的稳定性。数字司法实现价值对齐,乃是技术理性回归司法本位的核心路径。这一目标要求技术应用必须深植于司法的价值谱系,以规范算法逻辑、构建统一裁量参考系为抓手,将形式正义与实质正义的内在要求内嵌于技术规则之中。此举既能通过数据闭环防止裁量权滥用,确保裁判尺度统一,又能以技术透明消解暗箱操作,维护司法权威以提升司法公信力。因此,即便司法形态随技术迭代不断演进,价值对齐始终是平衡司法变革之“变”与法理坚守之“不变”的关键,确保数字司法在效能提升与价值实现之间行稳致远、达致双赢。

(33) 参见张劲松:《人是机器的尺度——论人工智能与人类主体性》,载《自然辩证法研究》2017年第1期,第49页。

(34) 参见刘福森:《价值、主体性与历史唯物主义》,载《哲学研究》1993年第5期,第46页。

(35) 参见接婧:《国际学术界对鲁棒性的研究》,载《系统工程学报》2005年第2期,第153页。

(36) See Peter Bernard Ladkin, *Robustness of Software*, Digital Evidence and Electronic Signature Law Review, Vol.17:15, p.15-24 (2020).

第二,通过数字司法价值对齐实现司法价值的可塑性。王泽鉴认为,各国法律多以一定之人的形象为规范基准。⁽³⁷⁾此言揭示了法律与人的深层共生关系。不同的群体形象塑造了特色各异的法律制度,进而沉淀出独特的司法价值内核。司法价值并非无源之水,其生成既植根于社会结构与理论体系的交融,其形塑亦依托于特定时代的群体画像。因此,司法价值绝非亘古不变,而是随时代演进不断迭代的动态体系,呈现出鲜明的时代特征。鉴于此,数字司法的构建须秉持辩证思维:既要坚守司法价值的本质属性以维护制度根基,又要顺应技术变革赋予其在特定时空与主题下的弹性与可塑性,最终确保其精准契合数字时代的治理需求,实现基础理论与现代技术的有机统一。

第三,通过数字司法价值对齐实现司法价值的适应性。随着数字司法的发展,算法应用场景从裁判辅助拓展至合法性判断甚至智能裁判等领域,场景变化可能带来价值位阶的模糊,从而模糊司法价值的本来意义。这种场景化的价值对齐,需要算法在功能拓展时始终以司法程序确定为基准,避免因场景复杂而出现程序简化甚至程序剥夺的风险,实现算法对场景变化的动态适应。因此,数字司法价值对齐始终要按照司法价值的要求,适应当事人程序权利保护的需要,遵循司法程序的公开性、中立性、救济性,避免算法因追求场景适配效率而稀释程序价值及其相应的实体价值。

4. 避免恣意性

司法既是约束权力恣意的重要控制手段,也是保障公民权利的关键制度安排。若司法权缺乏刚性法律约束,其裁判功能极易偏离权利保障与权力监督的初衷,甚至异化为司法擅断,滋生司法腐败。在数字司法场景中,恣意裁判的倾向更容易被放大,价值取向也可能偏离社会共识。因此,必须通过价值对齐,确保数字司法价值的统一性与共识性,防范选择性司法,守住司法公正底线。

第一,通过促进司法价值的共识性避免恣意性。数字司法运行中若缺乏有效约束,极易为追求技术目标而偏离司法本质。价值对齐需依托制度与技术双重路径,明确要求数字司法始终恪守司法核心价值,杜绝因地域差异、平台偏好、供应商选择等导致价值取向随意化。数字司法强调价值共识,并非否定价值多元性,而是坚守司法领域已形成普遍共识的核心价值,排除非司法价值的不当干扰。同时,排除外部干扰不等于司法封闭运行,而是防止非司法因素对裁判结果形成实质性影响。在此过程中,尤其要避免片面追求司法效率等功利目标,侵蚀算法设计的公正性与价值中立性。

第二,通过警惕数字司法价值的擅断性避免恣意性。数字司法价值对齐通过制度设计与技术优化,杜绝数字司法在该优先何种价值上自主选择,将其价值判断逻辑牢牢锁定在司法价值的框架内,最终实现不同场景、不同系统中数字司法价值标准的统一,避免因技术逻辑异化或地方差异导致司法价值碎片化。数字司法价值对齐防止数字司法权擅断的要求具体表现为如下两点:一是在数字司法场景中,无论是民事案件还是刑事案件,司法价值要能够得到无差别的体现;二是在数字司法系统中,无论是案件事实认定系

(37) 参见王泽鉴:《民法总则》,中国政法大学出版社2001年版,第35页。

统还是结果裁判系统,都应当始终贯彻司法价值。

第三,通过防范数字司法价值的选择性避免恣意性。确定性与秩序本身就是社会福利的一部分。司法价值的确定性也是人们一直讨论的话题,并一直在司法实践中努力推行。人类社会发展中,具体价值虽不断变迁,但对核心价值的追求始终未变。在算法实践中,若缺乏核心司法价值锚定,数据筛选偏好、技术依赖等问题易催生新型选择性司法,加剧恣意风险。数字司法中,价值对齐的初衷是通过制度设计、技术赋能与伦理塑造,避免算法实践中因数据筛选偏好、技术依赖而滋生的司法恣意。

四、数字司法价值对齐的过程机制

数字司法作为司法领域与数字科技深度融合的产物,其价值对齐绝非单纯的技术适配或制度调整,而是关乎司法本质、社会秩序、国家治理乃至全球司法协作的重要议题。按照对齐数字司法价值规范基准、落实价值对齐具体要求的思路,数字司法价值对齐需要在源头校准层、运行管控层、迭代优化层与安全保障层系统构建层次化推进的过程机制,推动价值对齐与司法本质、技术应用、实践需求深度融合,确保数字司法始终坚守核心价值导向、彰显司法本质属性。

(一) 筑牢数字司法价值对齐的算法根基

在数字司法价值对齐的推进过程中,源头校准层需要将司法价值嵌入数字司法技术研发的初始环节,从源头规避算法偏见、数据失真等价值偏离风险,夯实数字司法运行的算法基础。

1. 将司法价值嵌入算法逻辑

在数字司法价值对齐的实践中,将司法价值转化为可执行的算法逻辑是技术嵌入的核心环节,也是筑牢数字司法价值对齐的前置基础。技术层面的价值嵌入突破了价值理念悬浮于技术之上的困境,将抽象司法价值拆解为量化指标、规则模块与约束条件,让算法在运行中主动遵循司法的价值导向。

第一,价值具象化是司法价值转化为算法逻辑的基本前提。法律价值在信息化时代,不仅需要体现人类的情感和道德需要,也需要可计算。⁽³⁸⁾甚至还有学者认为,可计算性标准的兴起与动态规则的演进,终将成为这个时代对法律表达范式的历史性贡献。⁽³⁹⁾因此,要实现价值对齐,就要将价值进行具象化表达,使司法价值成为可量化、可判断、可校验的具体指标,形成“价值—指标”映射的体系。从数字司法的全流程场景来看,司法价值的具象化表达应当按照不同的场景进行拆解,形成针对性的价值预设指标体系。如在司法审判中,可将正当程序价值依托指标识别转化为具象价值。具体核心指标为:一是庭审参与完整度,要精准统计当事人发言时长占比、关键意见表达次数;二是异议响

(38) 参见彭中礼:《论数字司法中价值的数据化及其挑战应对》,载《行政法学研究》2025年第5期,第10页。

(39) See John O. McGinnis & Steven Wasick, *Law's Algorithm*, Florida Law Review, Vol.66:991, p.1049 (2014).

应时效与质量,要考核法官异议回应时长,评估回应的针对性与合法性;三是程序告知完备率,要对照法定告知清单,量化权利义务、庭审流程、法律后果告知的完整性与及时性;四是质证权利保障度,要围绕质证环节实质化,评判质证权实现成效,杜绝质证形式化;五是庭审流程合法率,要对标法定庭审规范,量化回避、举证、辩论、最后陈述等环节的合法性程度。依托上述量化可核验参数,可将抽象程序正义转化为算法要素,实现司法价值精准内嵌于算法设计。

第二,使具象化价值进行算法逻辑表达是司法价值转化的核心要义。具象化价值契合算法逻辑,强调司法价值的技术表达与算法能力的价值承接,是数字司法遵循司法规律的重要体现。具体而言,可以分为如下阶段执行:一是在理念构思层面,要预设能够被算法实现的价值具象化技术。这需要算法工程师从技术层面设计衡量指标或者校正指标以表明价值的具象化可行性。二是在具体落实层面,要让数据成为价值的承载体。“数据是一种资源……通过海量数据训练,人工智能迅速掌握技能,从而在很多领域辅助人类完成各项工作,实现人类生产力水平的提升。”⁽⁴⁰⁾所以,数据是一种绝佳的决策工具和价值传递载体。在数字司法全流程中,通过对数据的筛选、关联、计算、校验等使用行为,将抽象的司法价值转化为可落地的具体实践,让数据成为价值传递的介质。

第三,将法律价值特别是司法价值嵌入算法模型,实现价值导向与技术逻辑的有机融合。算法逻辑与法律规范的深度融合,核心是将算法运行的技术原理与法律的规范体系,通过系统性整合实现相互赋能。具体而言,可以从三个层面推进:一是对法律条文进行语义解构,从而将法律规则和基本法律原则嵌入算法模型。对于自然语言所表述的法律条文,算法语言无法直接识别,需推进“代码即法律”向“代码循法律”的转变。“人的指令先通过高级语言传达给某个应用层,再通过逐步翻译、解释变成硬件能直接反应的机器语言。”⁽⁴¹⁾把法律变成可供算法遵循的语言,核心是打破自然语言的模糊性、开放性与算法逻辑的精确性、封闭性之间的壁垒,通过语义解构实现法律规则的形式化、逻辑的编码化,让法律从纸面文本变为算法可识别、可运算、可追溯的规则。二是将法律逻辑转化为算法逻辑,实现法律规则向数字代码的形式化转化。语义解构后的法律要素需进一步转化为算法可运算的形式化语言,打破自然语言法律文本与机器可读代码逻辑的壁垒,将法律的规范要求、构成要件、裁量标准转化为可被算法执行的精准代码,既确保代码对法律规则的忠实还原,又实现法律适用的自动化与标准化。例如,可以针对法律规则的构成要件,按“法律规则=构成要件+法律后果+例外情形”的逻辑结构,对条文进行层级化拆解,形成意义明确的规则单元。三是将法律价值特别是司法价值运用算法规则表达出来。“基于模拟大脑运作的机制、深度学习理论以及计算机对人类心理状态的阅读,人工智能能够处理价值问题,并通过属性特征、属性值以及最终形成特征向量的过程使得价值数据化,形成价值数据集。”⁽⁴²⁾在数字司法中,价值、技术与规范应当深度耦

(40) 盘和林:《新一轮科技革命的核心是数据》,载《每日经济新闻》2023年4月27日,第6版。

(41) 邓玲婷:《论计算机中法律规则的通用表示方法》,载《数字法治》2023年第3期,第111页。

(42) 彭中礼:《司法人工智能中的价值判断》,载《四川大学学报(哲学社会科学版)》2021年第1期,第160页。

合,才能够契合司法发展的时代要求。通过设定权重参数、约束条件和优先级规则,在算法中体现公平、正义、比例原则等。如在刑事案件中运用司法自由裁量权时,不能将“刑期”简化为“犯罪金额×情节系数”的机械计算,而应保留对社会危害性、悔罪表现等非量化因素的权重设置,确保算法输出的结果符合司法价值的要求。

2. 为数字司法价值对齐算法培育高质量数据

算法的有效运行离不开数据的支撑。高质量数据是数字司法价值实现的基础,其真实性、完整性、公正性直接决定算法运行的科学性,为公正、高效、便民等价值落地提供坚实的数据保障。然而,数据质量参差不齐的问题,在科技领域早已成为共识。⁽⁴³⁾脏数据、错误数据更是无处不在,容易产生消极影响。⁽⁴⁴⁾在数字司法中,数据污染不仅可能直接导致算法决策偏差、证据认定失误,而且可能使司法裁决结果与公平正义背道而驰,降低司法公信力。培育高质量数据的相关制度包括以下方面:

第一,建立数据采集标准制度。要明确数字司法数据采集的范围、格式、编码规范与填写要求,统一案件信息、人员信息、证据材料等各类数据的采集标准,细化数据采集颗粒度,⁽⁴⁵⁾修订完善案件信息业务标准,适配审判执行各业务条线精细化分析需求,避免因标准不一导致的数据混乱、缺失问题,实现数据的规范性与统一性。

第二,建立数据质量保障制度。该制度应当由以下制度构成:一是司法数据标准统一制度,明确司法数据采集、加工和存储等的统一规范;二是全流程司法数据质量控制制度,覆盖数据采集、清洗、存储、应用的全生命周期;三是司法数据质量审核与核查制度,建立事前审核、事中监测、事后核查的系统性规则;四是司法数据分类分级与权益保障制度,依据数据来源(公共、企业、个人)与敏感程度分类分级管控,对涉案个人信息、涉密数据强化脱敏处理与隐私保护,明确数据各参与方权责,规范数据加工使用权限。

第三,建立数据清洗校准制度。该制度聚焦反向纠偏,通过驱除脏数据、修正错误数据,确保司法算法所依赖的数据始终保持纯正与可靠。一是从制度内容看,要以数字司法价值为标尺明确核心制度内容,优先划定歧视性数据清除标准,将年龄、性别、身份偏见等违背司法公正的数据列为必清项。同时,规范数据清洗全流程,明确清洗范围、操作步骤、结果校验要求,留存清洗过程完整记录。二是从清洗方法看,要整合多元清洗方法形成适配体系,分类规范各类方法的适用场景与操作准则。如基于完整性约束的方法用于排查数据缺失、格式异常问题;基于规则的方法适配歧视性数据精准筛查,通过预设司法公正规则定向清除偏见数据等。三是从责任主体看,明确多元主体责任边界,落实全链条追责。司法机关作为核心主体,统筹负责并牵头制定清洗标准,承担最终审核与责任兜底义务;技术服务提供商为操作主体,对操作合规性与结果准确性负责;数据提供方承担源头责任,确保原始数据无刻意偏见信息。

(43) 参见郭志懋、周傲英:《数据质量和数据清洗研究综述》,载《软件学报》2002年第11期,第2076-2082页。

(44) 参见郝爽等:《结构化数据清洗技术综述》,载《清华大学学报(自然科学版)》2018年第12期,第1037页。

(45) 有媒体报道称,颗粒度更细的数据集能改善AI幻觉问题。参见徐贝贝:《更专业、颗粒度更细的数据集能改善AI幻觉问题》,载《金融时报》2025年3月28日,第2版。

（二）强化数字司法价值对齐的流程约束

在数字司法价值对齐的推进过程中,运行管控层的核心旨趣在于,以常态化制度与动态审查为抓手,确保数字司法技术在应用全流程中能够及时识别并纠正价值偏差,从而筑牢价值对齐的持续性与稳定性防线。这一目标的实现,有赖于一系列制度、机制作为支撑。

第一,建立预设和校准数字司法价值模型的制度。计算机编程形成的模型预设了法律数据的“输入—输出”方式,为数字司法价值对齐算法模型提供基础预设模式。具体而言,该制度包括:一是从制度内容看,以数字司法价值为核心明确制度核心内容,划定模型预设校准的范围、标准与底线。聚焦数字司法价值,细化算法参数设定、逻辑框架、决策阈值的校准规范,严禁嵌入歧视性、倾向性预设。二是从校准方法看,采用分层校准、双重核验并融合动态迭代推进方法。分层校准即先完成核心参数价值适配校准,再开展场景化细节校准;双重核验结合算法自动校验与跨领域专家评审,邀请各领域专家联合核查校准结果。三是从校准类型看,在数字司法算法编码开发阶段,针对法律规范与法律原则两类核心司法要素,分别构建专项测试机制与动态校正制度,规避仅依赖法律规范“单刀直入”的机械适用局限,实现技术赋能与司法规律深度契合。四是从责任主体看,要遵循权责对等、全程追溯原则划分多元主体责任。司法机关作为主导主体,牵头制定校准标准,审核校准结果,承担最终责任兜底义务;技术研发主体负直接责任,严格按照标准开展预设校准,保障算法逻辑与参数设定合规;第三方伦理审查机构承担监督责任,独立核查校准过程的价值适配性。

第二,建立数字司法价值对齐算法的审查和备案制度。《最高人民法院关于规范和加强人工智能司法应用的意见》中就提到要“坚持技术研发、产品应用、服务运行的透明性,保障人工智能系统中的……各个环节能够以可解释、可测试、可验证的方式接受相关责任主体的审查、评估和备案。”数字司法在建设过程当中,不仅应当将价值对齐算法作充分论证,也需要通过程序上的审查和备案制度保证价值对齐算法符合司法价值的要求。价值对齐算法开发者在应用该算法之前应当将算法模型方案(含校准标准、技术路径)报有权部门审查和备案,相关材料需包含价值偏差应急预案,并明确未来校准的触发条件。其具体流程包括四个步骤:第一步由算法研发主体提交备案材料,含算法逻辑、价值校准方案、测试报告等核心文件;第二步是司法机关收件后,在特定工作日内完成材料完整性审核;第三步开展实质审查,联合技术、法学和伦理学专家,核查算法价值适配性与合规性,通过模拟案例测试验证效果,出具审查意见;第四步是将审查通过的算法予以备案。

第三,建立数字司法价值对齐算法的透明与公开制度。在人类社会,透明是交往顺畅的前提。⁽⁴⁶⁾在数字司法中,“算法透明原则,既适用于政府的算法规制,也适用于私

(46) See Meg Leta Jones, *The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood*, *Social Studies of Science*, Vol.47:216, p.216-239 (2017).

营机构的算法规制”⁽⁴⁷⁾。算法透明要求算法的可解释性与司法公开原则相结合,让技术决策过程成为司法透明的有机组成部分。算法的核心逻辑需向司法参与者公开,这是破解司法算法黑箱、保障司法透明与程序正义的关键举措。具体而言,可结合不同司法参与者的角色定位与实际需求,构建分层、精准、可理解的公开体系。其一,对司法参与者而言,要消除数字鸿沟,保障程序参与权、知情权。当事人、律师等是司法程序的直接参与者,其主体性的丧失往往源于自身的数字能力不足或被技术流程排斥。为此,需要通过确立技术适配、线下兜底等制度,确保所有参与者能平等享受数字司法的便利,而非因技术导致边缘化。其二,对司法监督者而言,要打破技术黑箱,保障知情权与监督权,确保司法监督者的主体性。公众对司法的信任源于看得见的正义,数字司法的技术复杂性若转化为“黑箱”,会削弱公众的监督能力。⁽⁴⁸⁾ 监督是数字司法获得更好发展的有效途径。确保数字司法系统透明和负责任,将有助于建立对这些系统的信任和信心,并确保它们得到合乎道德且有效的使用,⁽⁴⁹⁾ 推动数字司法工具与司法规律、公众需求深度适配。

第四,建立数字司法价值对齐算法的第三方评估制度。作为一种治理工具,第三方评估具有广泛的制度意义。在算法深度应用的行业领域,第三方评估制度亦逐步兴起。此类评估能够为算法使用者提供客观依据,进而推动开发者持续优化算法体系。第三方评估具备独立性、专业性、权威性鲜明特征,评估结果相对公正、客观。⁽⁵⁰⁾ 通过建立第三方评估制度,客观验证算法价值主张与利益相关方诉求的一致性,可以防范算法偏差引发的伦理风险、权益损害及社会争议。其具体制度内容包括:一是从评估主体看,明确评估主体为独立第三方机构,需与司法机关、算法研发主体无利益关联,保障评估中立性。二是从评估资质看,设定严格资质准入标准,第三方机构需取得司法行政部门备案许可与数据安全等级保护认证,具备固定办公场所与专业团队。三是从评估能力看,强化评估能力建设,第三方机构需搭建适配数字司法场景的评估体系,具备算法逻辑解析、数据质量校验、价值偏差识别等能力。四是从评估效力看,明确评估结果的法定效力与应用场景,评估报告作为算法备案、迭代优化、责任认定的重要依据,司法机关需将其纳入算法监管档案。

(三) 推动数字司法价值对齐的迭代优化

在数字司法价值对齐的推进过程中,迭代优化层的核心目标是实现价值对齐的优化升级,通过全流程追溯、问题整改与经验转化和标准更新,推动制度与技术持续适配司法价值需求。

第一,建立数字司法价值对齐算法迭代的可追溯制度。算法在任何程序中的使用

(47) 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22页。

(48) See Gyandeeep Chaudhary, *Unveiling the Black Box: Bringing Algorithmic Transparency to AI*, Masaryk University Journal of Law and Technology, Vol.18:93, p.93-124 (2024).

(49) See *Id.*, 93.

(50) 参见徐双敏:《政府绩效管理中的“第三方评估”模式及其完善》,载《中国行政管理》2011年第1期,第28页。

都并非无形无迹,而是会留下可追溯的操作痕迹。⁽⁵¹⁾从数据输入、参数运算到结果输出,算法运行的每一个环节都会在系统中生成相应记录。这些记录既包括数据流转轨迹、运算逻辑节点,也涵盖结果生成的时间与条件,为后续的算法监管、问题排查和责任追溯提供了基础依据。算法溯源更多关注的是算法软件在程序实现过程中输出量值的准确可靠,⁽⁵²⁾因此,算法的可追溯性应贯穿算法全生命周期。一是从概念看,算法可追溯性是对算法设计、开发、运行等过程的关键节点进行结构化记录、存储与关联,保持算法全生命周期的可记忆性,确保任何环节的算法操作、数据或决策都可被反向查询、验证与还原。二是从方法看,可追溯制度要构建技术留痕与合法性审查相结合的双重追溯体系。在技术层面,依托区块链存证等技术,对算法训练数据集、模型结构调整等关键信息进行哈希加密与链式存储,⁽⁵³⁾实现全流程不可篡改;在法律层面,建立“版本回溯—差异比对—价值校验”的审查方法,通过对比迭代前后算法在相似案件中的决策输出,校验其是否偏离预设的司法价值。

第二,建立数字司法价值对齐算法问题导向的持续改进制度。该制度聚焦数字司法实践中价值对齐的现实困境与显性问题,包括以下四个层面:一是在问题层面,构建多源联动的线索采集机制,确保数字司法价值偏差问题被全面、及时识别。既要打通算法运行的全场景数据接口,自动抓取案件裁判中的算法决策输出等结构化数据,标注价值冲突的异常节点,也要建立常态化的人工排查渠道,定期梳理典型案例,明确问题描述、提交主体,为后续研判提供清晰、可追溯的基础素材。二是在研判层面,需依托技术和法律融合的双维度分析框架,定位问题产生的核心诱因。在技术上,组织算法研发人员、数据工程师对模型架构、训练数据集、参数设置进行全链路复盘,排查是否存在数据样本偏差、算法逻辑漏洞等技术缺陷;在法律上,由法学专家、资深法官对照司法价值,分析算法设计是否违背法律、法规、司法解释等。三是在整改层面,需针对根源研判结论制定差异化的修复策略。针对数据样本偏差问题,扩充多样化、代表性的案件数据集,剔除带有偏见的样本,补充特殊类型案件的标注信息,优化训练数据的价值导向性。四是在优化层面,推动算法与司法价值的动态适配。在整改方案试点验证通过后,将优化后的算法模型纳入正式运行体系,建立版本迭代备案制度,明确新版本的价值锚点、优化内容与适用范围。

第三,建立数字司法价值对齐算法的经验转化与标准更新制度。数字司法实践中,算法可能出现价值偏差、与司法需求脱节等问题,而经验转化与标准更新正是破解这一难题的重要支撑。推进算法价值对齐的有效经验、成熟做法要进行提炼、固化,同时要动态更新算法设计、数据应用、流程规范等相关标准,确保算法始终与司法本质、社会伦理、

(51) See Biagio Aragona & Francesco Amato, *Retracing Algorithms: How Digital Social Research Methods Can Track Algorithmic Functioning*, in Francesca Comunello, Fabrizio Martire & Lorenzo Sabetta eds., *What People Leave Behind: Marks, Traces, Footprints and their Relevance to Knowledge Society*, Springer, 2022, p.129-140.

(52) 参见王亭亭、崔伟群:《算法溯源简述》,载《计量科学与技术》2023年第5期,第26页。

(53) 参见慎君:《“只能加密不能解密”的哈希算法是什么》,载《保密工作》2023年第5期,第49页。

公众需求保持一致。该制度的核心框架为：一是在制度内容上，需以经验沉淀、标准迭代和价值锚定为核心框架，明确制度适用范围、主体权责与运行规范。二是在转化路径上，要融合实践、理论与技术三个维度，实现司法经验向算法标准的精准转化。从实践维度看，建立典型案例萃取机制，从算法价值对齐的成功判例中提炼裁判规则、价值导向与适用场景，形成可复制的实践经验库；从理论维度看，组织法学与算法专家对实践经验进行解构，将隐性的司法价值判断转化为显性的理论模型与量化指标；从技术维度看，推动理论指标向算法代码与参数的转化，嵌入算法训练与决策流程，确保经验转化的技术可行性与价值一致性。三是在联动机制上，需构建跨主体、跨环节与跨层级融合的协同体系，保障经验转化与标准更新的高效运转。在跨主体联动方面，建立司法实务部门、算法研发机构、标准化组织、第三方评估机构的联席会议制度，定期通报经验采集进展、转化难点与标准实施效果。在跨环节联动方面，设计“经验采集→理论转化→技术适配→标准制定→落地应用→效果反馈”的流程，明确各环节的衔接节点与反馈机制。在跨层级联动方面，打通上下级司法机关的经验共享通道，提升制度的系统性与适用性。

（四）构建数字司法价值对齐的保障机制

在数字科技深度嵌入数字司法的决策场景中，安全保障层的闭环监管已成为维护人类主体核心地位的关键防线。其核心目标是在 AI 等智能工具提供司法决策支持的同时，始终将人类置于目标定义者、逻辑审查者、最终决策者的核心位置，避免因技术依赖导致人类主体地位被弱化、边缘化甚至替代。

第一，要通过公众参与吸纳多元价值诉求。司法价值的内涵具有社会性，若仅由算法开发者或企业定义价值，易导致企业利益优先而忽视公共利益。价值对齐需通过多元主体参与，形成符合社会共识的公平价值标准。“作为一套制度系统的公众参与，不论是在国家宏观的政治生活中，还是在微观的行政过程中，都被理解为健全国家民主制度、提升公共生活民主性和公共性的重要途径。”⁽⁵⁴⁾ 数字司法虽依托技术提升了效率，但其核心仍需锚定司法为民的本质。公众作为司法服务的直接受众与社会价值的承载者，其多元诉求是数字司法避免技术异化、专业封闭的重要校准器。唯有通过制度化的公众参与，将不同群体的价值期待嵌入数字司法的设计、应用与优化全流程，才能让数字工具真正对齐社会共识性正义。此外，公众参与机制的有效性还需破解认知鸿沟与代表性不足两大难题：一方面，需通过数字司法科普降低参与门槛，用典型案例说明公众诉求如何影响系统优化，避免因公众对技术或法律的认知不足导致诉求偏离实质；另一方面，需通过分层抽样确保公众参与的代表性，可通过基层自治组织协助收集诉求、组织线下会议等形式，避免参与权集中于数字素养较高的群体。

第二，要通过风险干预精准纠偏，防范人类主体地位丧失。无论是数字算法还是人类算法，每种算法都有自己的优缺点。⁽⁵⁵⁾ 因此，如何规避缺陷而放大优势，需要有必要的

(54) 王锡梓：《公众参与：参与式民主的理论想象及制度实践》，载《政治与法律》2008年第6期，第8页。

(55) See Cary Coglianese & Alicia Lai, *Algorithm vs. Algorithm*, Duke Law Journal, Vol.72:1281, p.1287 (2022).

监管措施。价值对齐的核心风险在于 AI 自主决策可能偏离人类需求,算法黑箱引发决策失控、价值偏差持续放大侵蚀主导权,因此制度设计需聚焦关键决策环节,通过强制介入确保人类的主体地位不被架空。风险干预制度正是通过提前预警、实时阻断、事后修复的全链条设计,将这些隐性风险转化为可控变量,从源头守住人类在技术体系中的核心位置。一是从风险预警来看,需建立多维度风险监测指标,精准识别主体地位弱化的信号。在数字司法领域,针对智能量刑辅助系统,可设置法官对 AI 建议的采纳率阈值、个案人工修正率过低等指标,防止法官因过度依赖 AI 丧失独立研判能力。二是在风险阻断环节,需明确分级处置机制,避免风险扩散。对偶尔出现价值偏差的轻度风险进行人工干预,要求技术方在规定时限内结合人类专家意见优化算法。三是事后修复环节需聚焦风险影响的挽回与制度漏洞的补位。一方面,对因 AI 风险受损的人类权益进行救济,建立 AI 失误补偿基金制度,提供兜底保障;另一方面,针对风险暴露的制度缺陷及时修订,不断补充完善指标体系,避免同类风险重复发生。

第三,要保留司法人员对算法结论的最终审核权,实现权利保障。在社会交往中,人类努力以主体身份出现,避免沦为他人支配或操纵的对象。因此,保留司法人员对算法结论的最终审核权,并非对技术的否定,而是对司法本质与价值对齐目标的坚守。这种不可或缺的最终审核权,其必要性根植于两个层面:一是从算法的固有局限看,司法人员的最终审核权是弥补技术缺陷、规避价值偏离的必然要求。算法的决策逻辑始终依赖训练数据与预设模型,其自身存在三重局限:训练数据偏差易导致决策结论失准;算法模型无法覆盖司法实践中诸多非量化的价值维度;法律原则的灵活适用、个案中的利益权衡,均超出了算法的机械运算能力。这些固有局限,迫切需要司法人员通过最终审核予以弥补和修正。二是从司法的本质属性看,最终审核权是维护司法权人类主导属性、保障价值对齐方向的核心锚点。数字司法价值对齐是让技术服务于司法为民的目标,而非让技术逻辑主导司法决策。若放弃司法人员的最终审核权,将算法结论直接作为裁判依据,实则是将司法权的核心价值判断与正义裁量让渡给技术,会导致司法异化为数据驱动的机械流程,偏离以人为本的价值初衷。此外,最终审核权还需通过制度细化实现从原则到操作的转化,避免沦为形式。一方面,需明确审核权的实质审查义务,要求司法人员不仅对算法结论的合法性进行审查,更要对合理性进行研判,且需在裁判文书中书面说明采纳或修正算法结论的理由;另一方面,需配套算法可解释性支撑,要求算法提供结论生成依据。如此一来,在数字司法价值对齐的推进中,司法人员对算法结论的最终审核权将成为技术工具性与司法主体性的平衡支点、效率提升与价值坚守的结合纽带,最终让技术真正成为赋能司法、守护正义的工具,而非主导司法、偏离价值的力量。

五、结 语

数字司法价值对齐并非技术功能与司法制度的简单耦合,而是以司法本质属性坚守为核心导向,通过多维度协同机制实现技术理性与司法价值理性动态均衡的系统性工

程。本文探讨数字司法价值对齐的逻辑机理、规范要求与过程机制等核心议题,不仅回应了数字司法价值对齐如何落地的问题,更厘清了技术赋能与司法本真的辩证关系。价值对齐的革命性意义在于,它打破了传统人机关系中技术主导、人类被动的单向桎梏。从技术路径来看,价值对齐机制通过将人类多元价值转化为技术可理解、可执行的逻辑,为人类赋予了主导技术方向、优化技术路径的主动权,让人机协同从效率适配的工具层面,跃升为价值共生的生态层面,最终推动技术进步与人类发展形成深度嵌合的良性循环。

需说明的是,本文聚焦于宏观层面的理论建构与框架设计,后续研究还可从两方面深化拓展:一是开展特定司法场景的实证研究,细化不同场景下价值权重的动态配置标准与适配机制;二是对数字司法特殊价值的实现路径进行深入探究。唯有持续推进理论深化与实践探索,才能使价值对齐机制与数字司法的发展演进形成良性适配,为司法现代化进程提供更为坚实的理论支撑与实践指引。

Abstract: Deviations between algorithmic decision-making and judicial values in digital justice can easily give rise to ethical and legal risks. To advance value alignment in the digital justice domain, it is essential to uphold the essence of justice to orient its development, prevent and control technical risks to consolidate its foundation, safeguard judicial trust to demonstrate its effectiveness, and integrate judicial civilization to elevate its practice. Value alignment in digital justice requires the value pursuit of algorithms to form dynamic adaptation with legal spirit, social ethics, and humanistic care. Its core requirements are maintaining consistency, ensuring controllability, preserving robustness, and avoiding arbitrariness. To promote value alignment, it is necessary to gather collaborative forces across the entire chain and process, and establish a well-structured, hierarchical promotion mechanism covering the source calibration level, operation control level, iterative optimization level, and security guarantee level. Theoretical analysis indicates that value alignment in digital justice transforms diverse human values into the logic that is technically understandable and executable, enabling the collaboration between humans and machines to leap from the instrumental level of efficiency adaptation to the ecological level of value coexistence. Ultimately, it fosters a positive cycle where technological progress is deeply integrated with human development.

(责任编辑: 赵毅宇)