

doi: 10.15936/j.cnki.1008-3758.2025.02.011

# 正当程序视角下人工智能辅助量刑的挑战与重塑

洪涛

(浙江大学 光华法学院, 浙江 杭州 310008)

**摘要:** 人工智能辅助量刑是智慧司法建设的重点工程,但在正当程序视角下可发现其应用带来了多重挑战,如自动化决策架空程序参与原则,歧视性决策有违程序中立原则等。事实上,以上挑战不仅源于人工智能的不确定性,更深层的原因是传统程序正义理论未将数字空间纳入诉讼场域,忽略对技术研发者的规制以及程序公开未延至可解释的时代局限。程序正义范畴下的技术性程序正义与传统程序正义均以尊严理论和司法公信力理论为根基,且前者在适用场域、规制对象、构成要素等方面可补足后者的时代局限,亟须“补充性”引入。技术性程序正义理论用于司法量刑场景时,其要素包括第一层的“以人为本”原则以及第二层的量刑系统的合规性义务和利害关系人的程序性权利。在此指导下,应当增设程序选择制度,明确算法披露制度,健全鉴定制度和专家辅助人制度以及确立智能问责制度,并拓展协同研发机制、算法备案机制、算法听证机制和司法培训机制作为联动。

**关键词:** 正当程序; 人工智能; 辅助量刑; 技术性程序正义

**中图分类号:** D 915.3

**文献标志码:** A

**文章编号:** 1008-3758(2025)02-0111-10

## Challenges and Reshaping of AI-assisted Sentencing from the Perspective of Due Process

HONG Tao

(Guanghua Law School, Zhejiang University, Hangzhou 310008, China)

**Abstract:** AI-assisted sentencing is a key project in the construction of intelligent justice, but from the perspective of due process, it has been found that its application poses multiple challenges, e. g., the principle of procedural participation being hollowed out by automated decision-making, and discriminatory decisions violating the principle of procedural neutrality, etc. In fact, these challenges don't merely stem from the uncertainty of artificial intelligence, but the deeper reason is that the traditional theory of procedural justice has not included the digital space into the litigation field, ignoring the regulation of technology developers and the limitation of procedural openness not advancing to the era of explainability. Under the category of procedural justice, the technical procedural justice and traditional procedural justice are both based on the theory of dignity and judicial credibility, and the former can supplement the latter's era limitations in terms of applicable fields, regulatory objects, constituent elements, etc., so “supplementarity” should be introduced. The theory under the judicial sentencing scenarios includes the first layer of the “people-oriented” principle, and the second

收稿日期: 2024-04-13

基金项目: 教育部哲学社会科学重大专项资助项目(2023JZDZ012)。

作者简介: 洪涛, 浙江大学博士研究生。

layer of the compliance obligations of the sentencing system and the procedural rights of stakeholders. Under this guidance, the procedural selection system should be added, the algorithm disclosure system should be clarified, the identification system and expert assistance system should be improved, and the intelligent accountability system should be established. Furthermore, the cooperative research and development mechanism, algorithm filing mechanism, algorithm hearing mechanism and judicial training mechanism should be expanded as linkage.

**Key words:** due process; artificial intelligence(AI); assisted sentencing; technical procedural justice

随着大数据、人工智能等数字技术的广泛应用,刑事司法领域出现数字化变革。因量刑活动直接涉及被追诉人人身权、自由权的限制,甚至生命权的剥夺,故人工智能辅助量刑成为重点工程。面对人工智能介入司法量刑的风险挑战,学界进行了讨论。从研究历程来看,大致经过三个阶段:第一阶段是人工智能用于司法量刑的可行性和正当性研究。有学者指出,司法判决讲究“情理法”一致,人工智能无法对人类非逻辑思维进行模拟,难以进行裁判说理<sup>[1]</sup>;也有学者认为人工智能通过学习既有案例中法官集体的量刑经验,能够实现科学的量刑预测<sup>[2]</sup>。随着智能量刑的实践反馈及研究的深入,人工智能可用于量刑的观点成为共识,第一阶段的研究宣告结束<sup>①</sup>。第二阶段是人工智能对法官主体地位的影响研究。有学者提出,数据前置性和算法依赖性的特征极易形成“数据主义司法观”,导致法官沦为可计算、可预测、可控制的客体<sup>[3]</sup>。但多数学者认为,当下人工智能是弱人工智能,难以胜任知识覆盖面大、技术含量高的司法工作,仅能发挥辅助作用<sup>[4]</sup>。尽管人工智能不会替代法官的结论在学理上成为主流,但由于知识鸿沟、算法黑箱等导致技术依赖的实践异化,故第二阶段的研究仍在继续。第三阶段是人工智能对正当程序的挑战与治理研究。部分学者认识到人工智能不仅对法官造成影响,更是对正当程序构成挑战,威胁着所有程序参与者(尤其是被追诉方)的合法权益。目前,有研究从刑事司法领域切入,指出人工智能对正当程序造成冲击,主张建构数字时代的程序正义理论<sup>[5]</sup>。虽然颇具理论启发,但因未结合司法量刑场景,难以就人工智能辅助量刑提供针对性理论供给。部分研究立

足司法量刑场景,但有的选择了一体化视角兼论实体和程序问题,致使程序方面阐释不够深入<sup>[6]</sup>。最新研究在分析人工智能辅助量刑风险的基础上,通过技术性程序正义理论的赋新对决策程序作了诉讼化改造<sup>[7]</sup>。然而可惜的是,一方面未区分人工智能辅助量刑的正当程序挑战与挑战形成的理论成因,将两者统归于风险,导致挑战分析不够系统及问题成因不够突出;另一方面提出的治理方案主要围绕庭审环节,内容较单薄且未认识到人工智能辅助量刑贯穿诉讼流程,缺少“溢出效应”的机制设计。鉴于此,笔者将首先系统分析人工智能辅助量刑的正当程序挑战,并进一步指出以上挑战与传统程序正义理论的因果关联。其次,论证技术性程序正义理论“补充性”引入的同时,对其进行场景化诠释以提供针对性理论供给。最后,在新理论的指导下,对人工智能辅助量刑进行制度重塑和机制拓展,希冀对量刑规范化改革和数字正义的实现有所助益。

## 一、人工智能辅助量刑对正当程序的多重挑战

关于人工智能辅助量刑对正当程序的挑战,首先要对“人工智能辅助量刑”与“正当程序”两个概念进行剖析。一般认为,人工智能辅助量刑是指运用大数据、云计算和图谱结构完成法律知识图谱建构,在此基础上借助自然语义技术进行类案识别和情节提取,并以人类反馈学习算法进行训练,最终实现量刑预测和偏离度测试<sup>[8]</sup>。正当程序是指具有一定合理性和正当性的程序<sup>②</sup>。依

① 例如,《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条明确规定“辅助审判原则”。

② “正当程序”和“程序正义”无本质差别,学界基本上是混用的。出于语境使用的考虑,本文将程序正义作为理论指导,正当程序作为应然目标。简言之,合乎程序正义的程序就是正当程序。

据判断标准分为程序工具主义和程序本位主义：前者认为正当程序主要以实体正义的实现为价值目标；后者则主张评价正当程序的价值标准是程序是否具备一些内在品质。随着法治建设的进步，尊重当事人的主体地位成为核心考量，评价标准倾向于独立的内在品质。因此，对人工智能辅助量刑的正当程序挑战命题之判断，应从程序内在品质的遵守情况论证。

### 1. 自动化决策架空程序参与原则

程序参与原则是程序公正的首要要素。该原则主张可能受到刑事裁判或诉讼终局直接影响的主体（以下简称“利害关系人”）应有充分的机会富有意义地参与裁判制作过程，并对结果发挥有效影响<sup>[9]</sup><sup>136</sup>。理解该原则时，可结合哈贝马斯的“交往行为理论”，即主体间的交往行为理性依赖于主体间的言语和行为互动，在承认彼此主体性地位的基础上达成沟通和共识<sup>[10]</sup>。简言之，利害关系人的始终“在场”和富有意义地参与，离不开与对方当事人、裁判者面对面的对话和辩论，借助语言和感官即时地陈述意见、表达主张。然而，这些设计在人工智能辅助量刑面前丧失了应有价值，无法确保利害关系人的主体尊严。具体而言，人工智能辅助量刑过程中利害关系人虽然在物理空间中“在场”，但智能量刑决策的场域在数字空间且瞬时杂糅完成，利害关系人受限于信息壁垒而无法参与到决策制作过程之中，实际上处于“缺席”状态。此外，人工智能辅助量刑基于的案例样本、算法模型及评估权重通常为“黑箱”笼罩且带有专业属性，利害关系人因知识鸿沟而无法进行质询，遑论有效影响裁判结果<sup>[11]</sup>。概言之，人工智能辅助量刑背景下利害关系人与对方当事人、裁判者在物理空间进行的言辞、感官上的交往行为之程序参与价值大打折扣，难以通过“过程控制”进而有效影响裁判结果。更令人担忧的是，对技术理性的人工智能进行广泛应用，可能使人们不自觉地对其结论产生敬畏感并予以认可，沦为智能司法的旁观者或附庸者。

### 2. 歧视性决策有违程序中立原则

程序中立原则是正当程序的基石。该原则是指，裁判者应在发生争端的诉讼各方之间保持一种超然和无偏袒的地位，确保任何一方在诉讼活

动受到平等对待和同等尊重<sup>[9]</sup><sup>138</sup>。学界主流观点认为，人工智能在司法量刑中仅是辅助作用，并没有对司法公正、司法中立等价值形成根本性挑战<sup>[12]</sup>。这一论断主要源于理论分析和逻辑推理，但事实上，人工智能辅助量刑至少在两方面对程序中立原则构成挑战：其一，对具有“追诉倾向”的智能量刑建议的直接采纳造成“控审合一”。数字检察背景下检察机关越来越多地依赖内部智能系统，但此类系统以其过往案件为训练数据，且算法编码时不自觉地融入追诉立场，导致智能量刑建议带有“追诉倾向”。实践中，即便法官对智能量刑建议的合理性存疑，也通常因专业知识的缺乏和技术分析的强逻辑感而选择信赖<sup>①</sup>。对具有“追诉倾向”的量刑建议的直接采纳，导致控审分离沦为空谈。其二，智慧法院建设中的量刑系统不可避免地因数据、算法而产生歧视，一旦法官未能及时发现便会破坏程序中立原则。一方面，量刑系统训练用的司法案例因样本天然偏差、大数据的拓展属性，在“垃圾进，垃圾出”的定律下极易导致歧视性结果；另一方面，技术人员选择算法变量时，也会因社会文化、价值偏好等因素不可避免地构成“既存性偏见”造成结果偏差<sup>[13]</sup>。美国法院使用的 COMPAS 系统将性别、种族等变量用于犯罪可能性评估，导致黑人被告的再犯风险概率远远高于白人被告，通常会受到更重的量刑惩处<sup>[14]</sup>。

### 3. 黑箱化运行减损程序理性原则

程序理性原则是正当程序的必备要素。该原则要求裁判者制作裁判的程序必须符合一定理性，确保其判断和结论基于可靠和明确的认知<sup>[15]</sup>。结合前文可知，人工智能辅助量刑依托“司法大数据”和“量刑算法决策”，并以“数据集中处理—意见智能生成”为运行范式：法律知识图谱建构、类案识别和情节提取是对数据的集中处理；模型训练通过挖掘量刑数据与量刑规律的相关关系，进而生成意见用于量刑预测和偏离度测试<sup>[7]</sup>。上述流程在实践中因黑箱效应而引发诸多质疑，一定程度上违背了程序理性原则。首先，量刑系统的训练数据来源不透明，包括但不限于地区覆盖范围、案件类型、时间跨度等信息，同时法律知识图谱的构建路径亦未明确，加上类案识别和情

① 数据显示，2021年全国检察机关运用量刑辅助工具提出的量刑建议，法院的最终采纳率近97%。参见林平：《检察量刑辅助工具提出建议，法院采纳率近97%》，网址：[https://www.sohu.com/a/498929278\\_260616](https://www.sohu.com/a/498929278_260616)。



节提取难以确保无遗漏,导致数据集中处理阶段的合理性存疑,与程序理性原则的要求相抵牾。其次,量刑系统的算法透明度匮乏且不具备可解释性,不仅裁判者和诉讼当事人无法了解其原理,甚至技术研发者也存在认知局限。美国卢米斯案中,被告人向威斯康辛州最高法院提出上诉,认为法院过度依赖 COMPAS 系统侵犯了他的正当程序权利,相关算法和数据被作为商业秘密而未公开,无法进行有效质询<sup>①</sup>。域外研究者指出,COMPAS 等量刑系统的算法处于黑箱状态,其内部结构难以解释,刑事被告及其辩护律师面对量刑结果时无法提出质疑,严重侵蚀着程序正义理念<sup>[16]</sup>。

#### 4. 公权力垄断冲击程序对等原则

正当程序主要面向抽象意义上的人,是各方“势均力敌”下的形式对等,忽略了由财产、地位等差异造成的诉讼能力差距。于是,以“实质对等”为核心的程序对等原则被提出,主张诉讼中强大的一方应当承担一些特殊义务,能力较弱的一方则应当拥有一些必要特权,以便实现“平等武装”<sup>[9][139]</sup>。目前,量刑系统的研发存在公权力垄断的情况,即主要通过检察院、法院等司法机关自主研发或技术采购的方式进行,律师团体、法学院校、普通民众等社会力量的参与不足<sup>[17]</sup>。公权力垄断下的人工智能辅助量刑加重了控辩失衡,原本难以实现的“平等武装”更加遥不可及。首先,被追诉人因信息收集、数据分析等能力缺陷而未参与到量刑系统的使用过程,加重了自身对于数据来源、算法原理等内容的认知障碍,无法进行有效质疑和辩护。相比之下,公诉机关无论是自主研发还是技术采购时,均可将符合其价值倾向的案例作为训练数据,并在模型训练时融入其偏好,使输出结果更有利于己方。其次,控辩双方借助人工智能获得的量刑意见,在诉讼活动中并未得到同等对待,更不用谈对弱势方的倾向性保护。人民检察院运用人工智能得到的量刑建议不仅具有法定依据,且通常被法院直接采信为最终裁决。相比之下,多数被告人在司法实践中缺乏使用官方量刑系统的渠道,即便少数情况下借助小包公、法宝智判等私主体的系统得到有利结果,也极少被法院采信。综上,人工智能虽是一把诉讼利器,但并未改善处于弱势地位的被追诉方的境遇,反

而进一步强化控诉方的优势,无益于程序正义的实现。

## 二、成因探析:传统程序正义理论的时代局限

目前,学界现有研究仅看到人工智能对正当程序单方面的挑战,未深挖背后的根本成因,提出的解决方案多落脚于技术的审慎应用(如限于简单案件,且由法官最终决断)和程序的针对性优化(如公开量刑算法、提供量刑救济)<sup>[6]</sup>。该进路一方面导致人工智能的潜力不能充分释放,另一方面又无益于诸多措施之间的相互衔接,难以实现系统性治理。事实上,数字时代下传统程序正义理论的局限逐渐显现,无法提供完全的解释力。

### 1. 未将数字空间纳入诉讼场域

国内外学者普遍认为,程序正义理论来自英美法系,其渊源可追溯至自然正义。该理论历经岁月演变,而该过程中人们的生产活动和社会活动主要发生于物理空间,诉讼活动亦不例外。故此,传统程序正义理论的诉讼场域也是以物理空间为面向,比如诉讼当事人有权在裁判制作过程中始终“在场”,可以向裁判者提交证据、提出主张等。随着数字技术的发展,人类创造出独立于物理空间之外的数字空间,使得数字社会成为现实。数字社会中的诉讼活动亦发生变革,但这种变革非始于人工智能,在此之前就已经开始,电子数据入法便是典例。面对新兴数字技术带来的挑战,传统程序正义理论虽然作了针对性优化,但尚未认识到其所面向的诉讼场域过于狭窄,无法规制数字空间中的诉讼活动。具体而言,由于传统程序正义理论未将数字空间纳入诉讼场域,物理空间中的利害关系人缺乏正当依据介入数字空间中的活动,如无法知悉量刑系统选用了哪些数据、训练了何种算法等。从上海市高级人民法院的“刑事案件智能辅助办案系统”和广州中级人民法院的“智审辅助量刑裁决系统”的运行现状来看,并未明确赋予当事人就类案司法大数据进行审查的权利,其背后原因亦是缺少对诉讼当事人在数字空间内应有权利之规定<sup>[18]</sup>。进言之,即便量刑系统保持黑箱化运行,或者内含法官的既有偏见、控诉方的追诉倾向,也会因算法本身处于数字空间

① 参见: State vs Loomis, 881 N. W. 2d 749 (Wisc. 2016)。

而获得规制“豁免权”，一旦这些问题得不到妥当解决，程序参与原则、程序理性原则在人工智能辅助量刑面前自然失效。

## 2. 忽略了对技术研发者的规制

为增强司法裁判的可接受性，传统程序正义理论对不同诉讼主体的行为进行了相应规制，以使诉讼活动符合司法公正。例如，要求审理案件的法官保持中立立场，赋予被告人免于自证其罪的权利等。传统程序正义理论将控诉方、辩护方和审判方作为核心规制主体，而对证人、鉴定人、翻译人等其他诉讼参与人的关注相对较少。然而，人工智能辅助量刑背景下，技术开发人员、系统部署人员等技术研发者虽不在以上刑事诉讼主体的范围内，却通过数据收集与分析挖掘、算法训练与模型建构等活动对诉讼进程、裁判结果发挥着实质性的影响，理应遵守正当程序，但是，依照传统程序正义理论的规制范围，又无法对技术研发者进行规制，即便他们的行为对司法公正造成了挑战。具体而言，量刑系统的建设不但需要大量资金，更离不开技术专家，当下系统主要通过“司法人需求导向+技术研发执行”的模式构建。在此过程中，法学知识与技术知识塑造了不同的权力逻辑，当技术研发者运用技术手段理解和满足司法需求时，便可能出现话语冲突和价值偏差，造成量刑结果违背司法正义<sup>[19]</sup>。例如，技术研发者在进行数据训练时，往往不对样本进行“对”与“错”的价值判断，可能将一些错案或者裁判依据过时的案件作为样本，导致量刑结果偏离法律规定<sup>[20]</sup>。简言之，随着人工智能介入量刑，技术研发者成为实质参与者，他们影响着诉讼的过程和结果，但就实践情况来看，却游离于传统正当程序理论的规制范围外<sup>[5]</sup>。

## 3. 程序公开未延至可解释

程序公开一般不作为正当程序的独立原则，但不意味着正当程序无须程序公开。恰恰相反，之所以没有将程序公开作为独立原则，正因为它 是程序公正的基石性要素，渗透在程序参与、程序中立等原则之中，是原则之原则。受时代所限，传统程序正义理论下的程序公开以“信息可感知”为核心，但在人工智能辅助量刑的场景下，基于“信息可感知”的程序公开遭遇前所未有的挑战。首

先，司法数据的训练、算法模型的建构以及自动化决策过程均发生于数字空间，超越了物理空间的直观可见范畴，诉讼参与者要借助技术手段穿透空间屏障才可获取相关信息，这无疑提升了“信息可感知”的标准；其次，量刑算法具有一定价值属性，往往是研发企业的知识产权或商业秘密，相较于普通信息的公开难度更高；最后，算法本身具有复杂性和专业性，即使对算法原理、运行机制、源代码等信息进行公开，普通民众乃至裁判者也无法实质性理解，遑论行使参与权、异议权。申言之，以“信息可感知”为核心的程序公开在人工智能面前失效了，这背后是因为“信息可感知”不等于“信息可解释”：前者仅关注信息的可视性，后者则更强调信息的理解性。对于算法而言，可解释才是关键，公开只是通往前者的一个阶梯，甚至不是必要手段<sup>[21]</sup>。传统程序正义理论尚未认识到从“信息可感知”延至“信息可解释”的必要性，面对人工智能在刑事司法中的应用，原有理论框架在很大程度上变得力不从心。

# 三、可行路径：技术性程序正义理论的尝试引入

面对持续发展的刑事司法实践，传统理论难免出现滞后问题，但这并不意味着我们可以弃理论于不顾，转投至“技治主义”的怀抱。传统程序正义理论固然有其局限，但并未完全丧失内在价值，可作为新理论的“底版”。因此，应当立足本国国情以及刑事司法的特定场景，积极推动传统程序正义理论的时代发展。国际上，一种专门用于算法自动化决策的技术性程序正义理论（technological due process）被提出<sup>①</sup>，可为人工智能辅助量刑的有效治理提供一条可行路径。

## 1. “补充性”引入的正当性论证

技术性程序正义理论最早由美国学者席特伦提出，其产生与20世纪50年代以来自动化决策在行政领域的广泛应用紧密相连。随着自动化技术逐步渗透至资格审查、风险评估和政府治理等业务，“技术统治”“算法霸权”的担忧不断弥漫。席特伦指出，算法权力主导下的自动化行政对参与、中立等程序正义要素构成挑战，需要重塑程序

① 严格翻译应是“技术性正当程序”，考虑到我国学界的用语习惯，本文选择了“技术性程序正义理论”的表述，同时也为了与“统程序正义理论”保持协调。

性机制以便向行政相对人、利害关系人和公众赋能<sup>[22]</sup>。从内容上看,技术性程序正义理论对自动化系统提出区分“规则”和“标准”、确保透明性和可问责性、生成和保存轨迹三点要求,并主张赋予个人“获得有效通知”和“被听取意见”的程序性权利<sup>[22]</sup>。现有研究主要在此基础上作细化性、调整性研究,将技术性程序正义理论作为算法治理乃至数字治理的基础理论。具体到刑事司法领域,有学者指出传统正当程序理论对人工智能背景下司法运作的解释失灵,为化解程序正义风险,有必要引入技术性程序正义理论<sup>[23]</sup>;也有观点认为,在传统程序正义理论的基础上引入技术性程序正义理论,可更好地解释和指导数字时代的刑事司法<sup>[5]</sup>。总的来讲,在刑事司法领域引入技术性程序正义理论逐渐成为共识,但新理论应以何种角色引入以及此种引入的正当性论证暂付阙如。

在笔者看来,技术性程序正义与传统程序正义有共通之处也有各自区别,两者并非对立关系而是互补关系,前者可补足后者的时代局限。首先,技术性程序正义和传统程序正义均属于程序正义范畴。程序正义理论的容纳性、灵活性较强,其内涵会随着时代发展而发展。简言之,程序正义理论具有内容上的包容性和适用上的无限伸缩性<sup>[24]</sup>,随着传统时代到数字时代的变化,相应地发展出“血缘同脉”的传统程序正义和技术性程序正义。其次,技术性程序理论与传统程序正义理论均以尊严理论和司法公信力理论为根基,不但强调要尊重人作为主体的尊严,也希望通过正当程序使利害关系人和社会公众信任和尊重裁判。传统程序正义理论强调的程序参与、程序中立等内容,归根到底是为了尽可能地尊重人的主体尊严,提升司法裁判的可接受性。技术性程序正义理论亦是如此,无论是对技术透明、可问责的要求,还是赋予个人“获得有效通知”和“被听取意见”的程序性权利,都是为了一种“善”的内在价值,即尊重人的主体尊严<sup>[25]</sup>。

同时,技术性程序正义理论与传统程序正义理论在适用场域、规制对象、构成要素等方面有所差异,前者对后者有着补充作用。在适用场域方面,传统程序正义理论主要面向物理空间,缺少应对发生于数字空间的数据收集、量刑计算等活动的手段;技术性程序正义理论恰恰以数字空间为主要面向,关注人工智能等数字技术的研发与应用。在规制对象方面,传统程序正义理论针对的

是诉讼当事人,无论是权利赋予、义务要求还是责任认定均基于此,人工智能的出现导致人的主体尊严额外受到数据、算法的威胁以及追责的挑战;技术性程序正义理论将数据、算法及背后的技术研发者纳入规制视野,并以分布式道德责任规训人机交互,指出人类作为责任承担者的同时,也进一步分析机器与法律后果之间的因果,避免将责任推卸给技术<sup>[26]</sup>。在构成要素方面,传统程序正义理论忽略了数字技术介入司法这一不可避免的趋势及其挑战;技术性程序正义理论涵盖技术规范赋能与权利有效赋能两个方面<sup>[27]</sup>,释放数字技术潜能的同时又可进行有效治理。技术性程序正义理论也有自身局限,不能全面替代传统程序正义理论,如无法用于单纯物理空间中的诉讼活动。

## 2. “补充性”引入下的要素诠释

同一理论用于不同场景时,其构成要素难免要进行相应调整。前文提及的技术性程序正义理论之要素针对的是自动化行政,与司法量刑在适用对象、基本原理等方面存在本质差别,如自动化行政处理行政法律关系而司法量刑处理刑罚问题。“补充性”引入的技术性程序正义理论,还需结合司法量刑场景进行要素诠释。目前,有关研究大多停留于刑事司法领域(未细化至司法量刑场景),如有研究认为技术性程序正义作为对数字技术融入刑事司法的评判标准,包含排除偏见、充分参与、程序对等、程序合理、问责有效五项基本要素<sup>[5]</sup>。还有学者尽管聚焦于司法量刑场景,但重心放在“以人为本”的价值理念上,未就技术赋能、权利赋能的要素内涵进行展开<sup>[7]</sup>。在笔者看来,针对性理论供给要求采用法治系统工程的思维进行要素诠释,既要分析理论内部诸要素的关系,也应当明确各要素的具体内涵。

司法量刑场景下的技术性程序正义理论,在结构上分为两层:第一层是“以人为本”的价值理念,其是理论建设的根本出发点和落脚点;第二层包括量刑系统的合规性义务和利害关系人的程序性权利,前者旨在通过优化量刑系统来实现规范化赋能,后者则通过增强利害关系人的对抗能力来确保人机交互中的主体地位,两者相辅相成。从内容上看,“以人为本”要求人工智能始终以服务人类为目的,坚持以安全、值得信赖和负责任的方式设计、开发、部署和使用。在技术性程序正义理论中,“以人为本”统领着量刑系统的合规性义务和利害关系人的程序性权利。当下,“以人为



本”已成为国际社会认可的治理原则,某种程度具有了类似国际公约的效力。席特伦亦将“以人为本”作为技术性程序正义理论的赋能取向,指出该理论根本上是为了捍卫人的主体尊严,避免被机器主宰<sup>[22]</sup>。

其一,量刑系统的合规性要求。为消解人工智能辅助量刑造成的程序正义风险,量刑系统应遵循以下要求:①辅助非主导。人工智能赋能司法的内在逻辑仍是“人主机辅”,过度依赖人工智能会导致司法异化,人工智能只是辅助者而非主导者。②透明可解释。量刑系统的黑箱化导致利害关系人难以参与、算法歧视等问题,故量刑系统应满足透明度和可解释——前者要求披露源代码、输入数据、输出结果在内的技术要素,后者要求对技术要素进行合理说明<sup>[21]</sup>。③中立无偏见。无论是训练数据还是算法模型一旦蕴含歧视性因素,均会影响量刑系统的最终输出。为此,既要实施源头控制,对数据进行标准化和规范化审查,同时建立算法影响评估,确保从开发到使用的全生命周期监管。④损害可问责。法律责任是系统合规的守门员,没有责任则合规义务宛如无物。量刑系统应当做到损害可问责,亦即确保损害可以被精准追溯、有效控制和及时补救。

其二,利害关系人的程序性权利<sup>①</sup>。人工智能辅助量刑的规范化运行,不能仅靠技术层面的系统合规,还要赋予利害关系人程序性权利以落实“以人为本”原则。具体来讲:①信息获知权。人工智能辅助量刑因活动场域处于数字空间、量刑系统以黑箱运行等缘故,导致利害关系人无法及时获知信息。为此,应当赋予利害关系人信息获知权,将是否使用量刑系统、如何使用量刑系统以及如何处理智能量刑结果等信息进行告知。②充分参与权。人工智能辅助量刑下利害关系人缺乏有效参与的途径,有必要赋予其充分参与权,通过算法披露、算法听证等制度机制实质性参与并主导人机交互,避免技术异化。③算法解释权。面对量刑算法的复杂性和专业性,利害关系人因知识鸿沟而无法理解,也无法进行有效质疑。尽管可通过技术手段增强算法透明度,但这还不足够,应当赋予利害关系人算法解释权,允许其对量刑算法提出质疑并要求使用者作出解释<sup>[28]</sup>。

④算法救济权。利害关系人应当享有算法救济权,以便在人工智能辅助量刑造成或可能造成权利损害时,向有关部门申请损害救济或人工替代。

## 四、人工智能辅助量刑的制度重塑与机制拓展

纵观现有规范,可发现“两高三部”发布的《关于规范量刑程序若干问题的意见》仍未就人工智能用于司法量刑场景中的正当程序问题作出回应。最高人民法院发布的《关于规范和加强人工智能司法应用的意见》(简称《智能司法意见》)也仅是规定了安全合法、公平公正等基本原则,而未对诉讼各方的权利义务、程序步骤等内容进行规定。从这个意义上讲,应当以技术性程序正义理论为指导,尽快完善有关立法并用于风险治理,力争实现“理论指导立法—立法规范实践—实践反哺理论”的循环。由于立法完善绝非短时间内能够完成的任务,应当优先进行制度重塑和机制拓展:前者作为法律原则和法律规则的桥梁,体现着法律原则的价值取向,又指引着法律规则的具体内容;后者作为法律制度的溢出设计,可弥补制度因相对宏观而在微观层面的不足,并拓展解决途径助力融贯治理。

### 1. 制度重塑

承前述,司法量刑场景下的技术性程序正义理论遵循“以人为本”原则,涵盖量刑系统的合规性义务和利害关系人的程序性权利两方面内容,这些将在制度重塑过程中彼此融贯,共同致力于量刑规范化改革和数字正义的实现。

其一,增设程序选择制度。人工智能辅助量刑提高效率、加快类案同判的同时也带来正当程序挑战,对当事人的主体地位构成威胁,而诉讼的根基是司法公正,故量刑系统的使用应是可选择的。首先,因被告人是量刑裁判的承担者,应由其作为程序选择权的享有主体,即决定法院是否使用量刑系统。至于检察院使用人工智能是否需要被告人同意应具体分析:协商性司法中的量刑建议作为协商部分,理应考虑被告人的意见,对抗性司法中的量刑建议则无须被告人同意。其次,鉴

① 有观点认为,该部分内容可以被传统程序正义理论所包容,都是对诉讼当事人的权利赋能。事实上,传统程序正义理论的权利赋能因适用场域、规制对象的缘故,并不能涵盖技术性程序正义理论的权利赋能。例如,传统程序正义理论下的信息获知权主要是信息公开而延至信息可解释。

于我国量刑系统尚未完全成熟,难以应对案情复杂、情节众多的案件,可考虑将使用范围限定在轻罪案件。此类案件的证据、事实相对清晰,不会对量刑系统提出过高要求。再次,为保障被告人切实行使程序选择权,裁判者应告知量刑系统的使用及影响后果,确认被告人是否真诚自愿适用人工智能辅助量刑。最后,程序选择并不只是量刑系统的“进入”,还包括量刑系统的“退出”。自由包括积极的自由和消极的自由,程序选择本质上也是一种程序自由,程序选择权理应涵盖“进入”与“退出”。有观点认为,被告人享有量刑算法决策程序的自愿选择权,但该选择是不可逆的<sup>[29]</sup>。对此,笔者持反对意见。《智能司法意见》第5条规定:“各类用户有权选择是否利用司法人工智能提供的辅助,有权随时退出与人工智能产品和服务的交互。”为避免肆意退出造成效率低下和资源浪费,可对程序退出作出适当限制,如要求被告人提供一定的证据、理由。

其二,明确算法披露制度。算法黑箱和算法歧视的解决离不开算法披露,唯有打开黑箱并解释原理才能有效遏制歧视性量刑。对于该制度应注意以下几点:一是精准理解算法披露的内涵。部分观点将算法披露等同于算法公开,认为其本质上是一个信息公开和流动的过程,主要用于消解民众与技术研发者之间的“信息鸿沟”,即“可感知”<sup>[30]</sup>。事实上,算法披露还涉及算法原理、算法逻辑,乃至源代码的说明,即“可解释”。二是算法披露直接关乎利害关系人能否参与以及何种程度参与司法量刑,利害关系人理应有权申请启动。法院作为司法审判机关,有义务确保量刑裁判的正确性,亦可依职权启动。三是算法披露应当遵循必要性原则。有观点主张量刑算法具有公共属性,故其披露应当是原则性和全面性的<sup>[25]</sup>。笔者认为,量刑算法并非完全没有私主体性质(如司法机关与企业合作研发),且很多情况下诉讼各方对量刑算法并无异议或仅有部分异议,一律披露没有必要也不现实。为平衡算法公开与商业秘密保护,算法披露原则上针对争议部分,帕斯奎尔提出的“合格透明度”亦主张按照不同场景对披露范围进行选择限制<sup>[31]</sup>。四是量刑算法具有复杂性和专业性,需由责任人员作出解释以确保披露的

有效性。技术研发者负责训练量刑算法,理应对其“产品”承担说明义务,出庭解释算法的训练过程、运行机理。检察院和法院作为服务使用者,也要对其使用量刑系统的步骤、方法以及量刑结果的处理作出说明。

其三,健全鉴定制度和专家辅助人制度。诉讼各方尤其是被追诉方通常缺少专业知识,易因“心理强制力”将量刑系统的输出视为标准答案。公权力垄断又进一步恶化了被追诉方的不利境遇,为实现“平等武装”需借助鉴定制度和专家辅助人制度。《刑事诉讼法》第146条和第197条规定了鉴定制度和专家辅助人制度用于强化当事人,两者亦可用于人工智能辅助量刑<sup>①</sup>。具言之,训练数据的选取、算法模型的建构、智能量刑的原理等均属于专门性问题,解决这些问题离不开专门性知识,而鉴定人和专家辅助人正是用来弥补当事人的认知鸿沟,两者具有耦合性。此外,鉴定人和专家辅助人还可帮助利害关系人理解被披露的算法原理、算法逻辑等技术要素,落实算法透明可解释。鉴定制度和专家辅助人制度因出庭条件、诉讼费用等因素,在司法实践中尚有局限。故此,需要限制法官自由裁量权,规定一些法定出庭质证情形,如控辩双方对量刑算法有较大争议、技术研发者不能清晰解释算法等。

其四,确立智能问责制度。人机交互使得责任承担问题变得复杂,一旦不能明确具体的责任主体,一方面会使得技术研发者失去对量刑系统合规的主动追求,另一方面也会造成利害关系人算法救济权的求助无门。智能问责制度包括以下几点:首先,人工智能不是责任主体,应当在诉讼当事人、技术研发者、服务使用者等法律主体中寻找担责方。我国《新一代人工智能伦理规范》明确要求,坚持人类是最终责任主体。其次,为确保有效问责,量刑系统的研发者、使用者以及使用情况均应记录在案,并在裁判文书中加强对智能量刑的释法说理,以便回溯至问责点,进行精准追责。再次,考虑到量刑系统的过错认定有一定门槛,诉讼当事人的举证能力有限,故原则上适用过错推定原则,即由技术研发者、服务使用者证明自身无明显过错。最后,救济形式包括程序救济和实体

① 严格意义上的司法鉴定仅指四大类鉴定,即法医类、物证类、声像资料类和环境损害类。浙江自动化行政执法实践中,出现针对算法和系统进行鉴定的做法,其通过使用清晰、平白的语言解释算法和系统的原理打消了行政相对人的质疑。未来可增设“算法和系统鉴定”项目,由司法部登记造册统一管理。



救济两种类型。程序救济主要针对量刑系统的退出使用、启动上诉程序、二审发回重审等程序性权利,实体救济则是民事补偿、国家赔偿等实体性权利。

## 2. 机制拓展

以技术性程序正义理论为指导,对人工智能辅助量刑的制度进行重塑,可加快量刑规范化改革,但法律制度于微观层面的缺陷以及适用场域的限制,又使得正当程序的实现尚有一定距离,需要进行机制拓展进而与制度重塑实现协调联动。

其一,协同研发机制。当下量刑系统采用的是技术研发者为主导的单一研发机制:一方面主要通过检察院、法院等司法机关自主研发或技术采购的方式进行,律师团体、法学院校、普通民众等社会力量参与不足;另一方面司法人员仅负责提出需求,技术研发者通过数据训练和算法模型执行需求。单一研发机制造成或加剧了人工智能辅助量刑对正当程序的挑战,如律师团体、利害关系人的缺位导致有效参与落空等。社会治理经验告诉我们,多元主体的协同治理不但可增强决策民主性进而减少歧视,还能整合资源优势,提高治理效能<sup>[32]</sup>。对此,有观点可能质疑,倘若公诉机关和审判机关的量刑系统底层一致,两者的量刑结果将趋于相同,庭审实质化恐沦为空谈。事实上,唯有多元主体参与量刑系统研发,系统才可能在话语权力和各方利益的博弈之间,无限逼近价值中立和客观公正<sup>[33]</sup>。具言之,多元主体特别是被告人、辩护律师参与量刑系统的研发,可将辩方立场带入数据训练、模型建构,矫正单一研发机制下的追诉倾向<sup>[5]</sup>;多元主体在协同研发过程中,还可通过行为交往达成共识,确保人工智能始终向善发展,恪守“以人为本”原则。

其二,算法备案机制。算法披露、智能问责主要针对个案中的量刑算法,故制度重塑仅能做到个案的、事后的程序正义,难以从根源上解决人工智能辅助量刑带来的挑战,需要算法备案机制提供支持。我国人工智能监管的三部主要规范——《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》均确立了算法备案机制。有观点可能认为,目前的算法备案限于“具有舆论属性或者社会动员能力”,量刑算法未在此列,但笔者认为,在数字技术的迭代下人们无法一步到位规定所有算法,这一点可从备案范围随着规范出台而

不断扩大得到佐证。当下算法备案以告知作用为核心,唯有风险非常明显或损害发生后,才可进行事后溯源追责且惩罚力度有限<sup>[34]</sup>。为落实溯源治理,一方面有必要引入许可型备案,对算法的运行机制、评估报告等信息进行实质审查,唯有审查合格的算法才允许用于司法量刑场景且需定期检验;另一方面可建立动态调整的“负面清单”,将存在歧视风险的量刑算法、研发违规系统的研发者纳入“负面清单”,禁止相关算法的使用或一定时期内(甚至永久)不再采购其服务<sup>[6]</sup>。

其三,算法听证机制。量刑系统的使用实际上贯穿整个诉讼流程,包括审前的智能量刑建议、审中的智能量刑裁判、审后的智能量刑救济等,但算法披露制度主要作用于法庭环节,需设置专门的算法听证机制。听证程序深度契合“公正的外观”“可预测性、透明性、合理性”以及“参与性”等正当程序价值<sup>[35]</sup>,将其用于量刑系统的异议处理能提高司法裁决的可接受性。对于该机制应注意以下几点:首先,从利害关系人的信息获知权出发,各阶段主持听证的机关应当进行告知,使前者了解算法听证的规定、要求及后果,以便及时申请听证。其次,当事人均有权申请算法听证,表达自己的意见和主张,维护己方合法权益。作为决策制定者的一方,也可依职权启动算法听证。再次,鉴于量刑算法的专业性和复杂性,可在听证中引入专业组和社会组:专业组由法学专家和技术专家构成,分别从法律与技术两方面对量刑系统进行审查;社会组由社会民众和公益组织构成,负责考量智能量刑结果的社会影响。最后,赋予算法听证裁决相应的法律效力,防止后果缺少导致机制空转。倘若听证后,认为量刑系统存在歧视或错误,即异议成立,有关智能量刑意见不得作为司法裁决的依据。反之,则可作为司法裁决的直接依据。

其四,司法培训机制。正是由于司法人员不懂量刑算法的原理,才会在“心理强制力”的影响下直接采信智能量刑结果,导致技术统治司法的异化。在技术性程序正义理论下,重塑算法披露制度、鉴定制度和专家辅助人制度可一定程度上化解挑战,但并非治本之策。越来越多的学者主张通过培训或选录复合型司法人员,以应对数字时代的司法活动<sup>[36]</sup>。在笔者看来,确立司法培训机制不但可行而且必要。司法人员学习技术知识无须达到或超过技术专家的水平,而只要掌握相

关术语的概念、特征等基础知识,理解披露后的算法原理和机制,能够排除不合逻辑的意见即可。具言之,可聘请人工智能领域的专家研发适合司法人员的课程,并通过大练兵、业务竞赛等活动检验学习成果,培养一批具有丰富经验的复合型人才。然后,通过总结司法实践中的宝贵经验,形成全国性的智能辅助量刑裁判案例库。为避免司法培训机制成为额外负担,严禁硬性要求某一司法人员掌握多种算法,或者将培训经历作为年度绩效考核是否合格的指标。

参考文献:

[1] 赵艳红. 人工智能在刑事证明标准判断中的运用问题探讨[J]. 上海交通大学学报(哲学社会科学版), 2019, 27(1): 54-62.

[2] 吴雨豪. 量刑自由裁量权的边界: 集体经验、个人决策与偏差识别[J]. 法学研究, 2021, 43(6): 109-129.

[3] 郑戈. 算法的法律与法律的算法[J]. 中国法律评论, 2018, 5(2): 66-85.

[4] 吴习彧. 司法裁判人工智能化的可能性及问题[J]. 浙江社会科学, 2017, 33(4): 51-57.

[5] 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[J]. 华中科技大学学报(社会科学版), 2023, 37(2): 18-29.

[6] 程龙. 人工智能辅助量刑的问题与出路[J]. 西北大学学报(哲学社会科学版), 2021, 51(6): 163-174.

[7] 丰怡凯. 人工智能辅助量刑场景下的程序正义反思与重塑[J]. 现代法学, 2023, 45(6): 98-117.

[8] 王禄生. 司法大数据与人工智能开发的技术障碍[J]. 中国法律评论, 2018, 5(2): 46-53.

[9] 陈瑞华. 程序正义理论[M]. 北京: 商务印书馆, 2022: 136.

[10] 尤尔根·哈贝马斯. 在事实与规范之间——关于法律和民主治国论的商谈理论[M]. 童世骏, 译. 北京: 三联书店, 2003: 84.

[11] Shi J H. Artificial intelligence, algorithms and sentencing in Chinese criminal justice: problems and solutions[J]. Criminal Law Forum, 2022, 33(2): 121-148.

[12] 孙道萃. 人工智能辅助量刑的实践回视与理论供给[J]. 学术界, 2023, 38(3): 112-128.

[13] Friedman B, Nissenbaum H. Bias in computer systems[J]. ACM Transaction on Information Systems, 1996, 14(3): 330-347.

[14] Angwin J, Larson J, Mattu S, et al. Machine bias[EB/OL]. [2005-01-16]. [https://www. propublica. org/ article/ machine-bias-risk-assessments-in-criminal-sentencing](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

[15] 陈瑞华. 刑事程序的法理[M]. 北京: 商务印书馆, 2022: 233.

[16] Fortes P R B. Paths to digital justice: judicial robots, algorithmic decision-making, and due process[J]. Law and Artificial Intelligence in Asia, 2020, 7(3): 453-470.

[17] 王禄生. 智慧法院建设的中国经验及其路径优化[J]. 内蒙古社会科学, 2021, 42(1): 104-114.

[18] 李哲. 当司法进入“人工智能时代”[EB/OL]. (2017-07-12) [2024-01-09]. [https:// www. sohu. com/a/ 156544511\\_120702](https://www.sohu.com/a/156544511_120702).

[19] 王禄生. 大数据与人工智能司法应用的话语冲突及其理论解读[J]. 法学论坛, 2018, 33(5): 137-144.

[20] 王禄生. 论法律大数据“领域理论”的构建[J]. 中国法学, 2020, 37(2): 256-279.

[21] 沈伟伟. 算法透明原则的迷思——算法规制理论的批判[J]. 环球法律评论, 2019, 41(6): 20-39.

[22] Citron D K. Technological due process[J]. Washington University Law Review, 2008, 85(6): 1249-1313.

[23] 李训虎. 刑事司法人工智能的包容性规制[J]. 中国社会科学, 2021, 42(2): 42-62.

[24] 罗英. 数字技术风险程序规制的法理重述[J]. 法学评论, 2022, 40(5): 151-160.

[25] 郭春镇, 勇琪. 算法的程序正义[J]. 中国政法大学学报, 2023, 17(1): 164-180.

[26] Floridi L. Distributed morality in an information society[J]. Science and Engineering Ethics, 2013, 19(3): 727-743.

[27] Citron D K, Pasquale F. The scored society: due process for automated predictions[J]. Washington Law Review, 2014, 89(1): 1-33.

[28] 张凌寒. 权力之治: 人工智能时代的算法规制[M]. 上海: 上海人民出版社, 2021: 61.

[29] 张玉洁. 智能量刑算法的司法适用: 逻辑、难题与程序法回应[J]. 东方法学, 2021, 14(3): 187-200.

[30] 张恩典. 算法透明度的理论反思与制度建构[J]. 华中科技大学学报(社会科学版), 2023, 37(6): 29-40.

[31] Pasquale F. The black box society[M]. Cambridge: Harvard University Press, 2015: 142.

[32] 夏锦文. 共建共治共享的社会治理格局: 理论构建与实践探索[J]. 江苏社会科学, 2018, 39(3): 53-62.

[33] 王玉薇, 高鹏. 人工智能算法司法应用的发展隐忧及完善路径[J]. 学术交流, 2023(11): 69-83.

[34] 张吉豫. 论算法备案制度[J]. 东方法学, 2023, 16(2): 86-98.

[35] Crawford K, Schultz J. Big data and due process: toward a framework to redress predictive privacy harms[J]. Boston College Law Review, 2014, 55(1): 93-128.

[36] 胡铭. 电子数据在刑事证据体系中的定位与审查判断规则——基于网络假货犯罪案件裁判文书的分析[J]. 法学研究, 2019, 41(2): 172-187.