Doi:10. 11835/j. issn. 1008-5831. fx. 2020. 12. 003

欢迎按以下格式引用:甄航. 人工智能介入量刑机制:困境、定位与解构[J]. 重庆大学学报(社会科学版),2023(4):191-201. Doi:10. 11835/j. issn. 1008-5831. fx. . 2020. 12. 003.

Citation Format: ZHEN Hang. Artificial intelligence intervention in sentencing mechanism: dilemma, orientation and deconstruction [J]. Journal of Chongqing University (Social Science Edition),2023(4):191-201. Doi:10. 11835/ j. issn. 1008-5831. fx. 2020. 12. 003.

人工智能介入量刑机制:困境、定位与解构

甄 航

(西南政法大学 法学院,重庆 401120)

摘要:司法人工智能分为常规型人工智能与专业型人工智能,前者是将通用领域已经发展成熟的人工智能直接移植至司法领域而无需专门的算法更新,主要目的是将审判人员从繁重的"事务性"工作中解放出来,因而其也无法介入审判的核心内容;后者是诸如量刑辅助系统等专门为司法领域开发的介入审判实质内容的人工智能,其是司法人工智能的核心。当前司法人工智能的实践现状是常规型人工智能因其有坚实基础而卓有成效,但极为重要的专业型人工智能的开发与使用并不理想,主要原因是法学研究对专业型人工智能的研发理论供给不足,具体表现为"抽象有余而具象不足",其深层次原因在于法学专业知识与人工智能技术知识没有深度融合,即"懂技术的不懂法律,懂法律的不懂技术"。宏观层面,在智能爆炸不可预期的时空背景下,生命2.0阶段(文化阶段)或弱人工智能时代仍是当下及可预见未来所长期处于的阶段,故作为"工具"的量刑人工智能仍应定位于辅助量刑而非决定量刑,且基于量刑规范化改革的价值内涵,应更进一步地定位于规范性辅助而非参考性辅助,二者的区别是智能系统给出的阶段性量刑结论对法官的约束力大小。在微观层面,智能量刑系统的算法构建应以量刑逻辑主导算法逻辑为原则,以诸如量刑基准、不法刑等具有"共性"属性的阶段性量刑为作用领域而非其能力之外的终局性量刑结论(宣告刑);此外,为防止算法黑箱、算法歧视以及相关关系代替因果关系,须做到量刑人工智能的算法公开和阶段性量刑结论的可解释性。

关键词:人工智能;量刑规范化;智慧司法;量刑辅助;算法黑箱;算法歧视

中图分类号:D924. 13;TP18 **文献标志码**:A **文章编号**:1008-5831(2023)04-0191-11

近年来,在移动互联网、超级计算、大数据、传感网、脑科学等新理论的驱动下,具有深度学习、跨界融合等特征的人工智能技术呈指数爆炸式发展。2017年7月8日,国务院印发《新一代人工智

基金项目:中国博士后科学基金第71批面上资助(2022M712650);2022年重庆市教育委员会人文社会科学研究项目"智慧量刑的实践困境与理论破解研究"(22SKGH028);重庆市2019年研究生(博士生)科研创新项目(CYB19130)

作者简介:甄航,西南政法大学法学院,Email:zhenhanglmn@163. com。

能发展规划》(以下简称《规划》),标志着我国将发展人工智能技术提升到了国家战略层面。人工智能革命的到来已成为不可逆趋势,其正以汹涌之势介入社会生活的方方面面,司法领域也不例外。“人工智能法学需要探讨‘法治实践的智能化’和‘智能技术的法治化’这两大维度”[1],本文主要在前者框架下进行剖析。在顶层政策的大力推动下,各地法院快速地建立起具有类案推送、量刑辅助、偏离预警等功能的智慧法院系统。但与政策层面大力投入相比,司法实践层面对人工智能的应用却存在诸多问题。“事实上,对于中国司法系统斥巨资力推的大数据及法律人工智能技术在司法实践中的运用效果并不理想”[2]。故此,本文以麻省理工学院物理学教授迈克斯·泰格马克提出的生命更迭的三种形态为背景,以现存及可预见未来的人工智能技术发展实况为基础,分析人工智能在司法领域的体系性地位,并在此基础上对人工智能介入量刑机制进行精细化的具象剖析①。

一、问题导向:抽象有余而具象不足

自2017年以来,刑法学界兴起了关于人工智能空前热烈的讨论,主要涉及人工智能刑事责任能力、人工智能介入司法领域的风险及防范、人工智能辅助量刑的宏观展望等问题。总体而言,现阶段对人工智能的法学研究表现出抽象有余而具象不足的特征,学者大多对人工智能介入司法领域进行宏观而抽象的分析,而无法具象化。这与现阶段我国法学话语体系与技术话语体系没有深度融合有很大关系,也即“懂技术的不懂法律,懂法律的不懂技术”[3],这就使得在法学领域讨论技术问题犹如隔靴搔痒,无法直击痛点。本文此部分旨在以司法人工智能的中国处境为背景,剖析我国人工智能法学研究之困境。

(一)司法人工智能的中国处境

“在官方正式决策的推动下,全国范围内的各行各业开始主动‘拥抱大数据’、研发应用各种智能系统”[4]。2016年7月,中共中央办公厅、国务院办公厅印发《国家信息化发展战略纲要》,提出“建设‘智慧法院’,提高案件受理、审判、执行、监督等各环节信息化水平,推动执法司法信息公开,促进司法公平正义”。2017年7月,国务院印发《新一代人工智能发展规划》,提出三步走战略。2021年3月11,全国人民代表大会通过的《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》提出加强智慧法院建设的要求。在顶层政策大力推动下,“全国86%的法院已建成信息化程度较高的诉讼服务大厅,各类智能诉讼服务设备层出不穷”[5],例如最高人民法院的类案智能推送系统、上海市高级人民法院的智能辅助办案系统(206工程)、北京法院的睿法官智能研判系统、贵州省高级人民法院的法镜大数据系统等。总体而言,司法人工智能可以区分为两种类型:常规型人工智能和专业型人工智能②。

第一,常规型人工智能。常规型人工智能是指将通用领域的人工智能技术直接应用于司法领域而不需要专门的程序设计或者只需要简单的程序设计的司法人工智能。“智慧法院”中能够实现

①以几十年前的图灵测试为标准界定机器是否“智能”已经无法满足当下的需求,且“中文房间模型”已经论证了图灵测试的局限性。但到现在为止,学界对“人工智能”的概念并没有达成一致意见,故本文在较为广义上使用“人工智能”概念,即迈克斯·泰格马克教授对智能的定义:完成复杂目标的能力。

②部分人工智能学者将人工智能区为分通用人工智能和专用人工智能,前者是指可以完成任何认知任务,并且完成得至少和人类一样好的人工智能(目前并不存在);后者是指可以完成一个较狭义目标组(例如下棋和开车)的人工智能(目前已经存在)。本文以人工智能适用领域进行分类,而非其智能程度,因为就目前而言人工智能之间并没有质的区别,以智能程度作为划分标准对本文而言无意义。

语音转文字、案卷数据化等功能的人工智能可以归为此类。常规型人工智能主要有如下特征:(1)此类人工智能在通用领域已经发展得较为成熟。(2)此类人工智能从通用领域转向司法领域较为容易,无需法学专业知识的引入。因为其虽然被引入司法领域,但其功能性并没有发生质变。(3)此类人工智能主要解决司法领域的"事务性"工作,而无法介入审判的实质内容。在司法实践中,常规型人工智能发展态势较好,能够将法官从沉重的事务性工作中解放出来,提高诉讼效力,也能够切实缓解我国"案多人少"的现状。常规型人工智能之所以能得到如此好的效果,主要是因为在通用领域已有较为成熟的技术沉淀,将其直接引入司法领域并不需要大量额外的、专门的算法更新。

第二,专业型人工智能。专业型人工智能是指为司法领域所独有,需要引入法学专业知识的人工智能。"智慧法院"中的量刑辅助系统、偏离预警系统等可以归入此类。专业型人工智能具有如下特征:(1)为司法领域所独有,而并非由通用领域直接引入。算法设计以需求为导向,专业型人工智能所实现的需求在通用领域并不存在,这就使得其算法逻辑需要"另起炉灶",而不能直接沿用已存在的算法。(2)法学逻辑主导算法逻辑。专业型人工智能以解决司法实践问题为导向,法学逻辑是解决司法实践问题之根本,算法是法学逻辑智能化的工具。(3)专业人工智能较通用型人工智能发展较晚,技术不成熟。专业型人工智能并没有达到预期效果,例如,就"江苏智慧审判系统的应用情况来看,似乎也不太理想,江苏基层法院的部分法官甚至表示并未使用该系统"[2]。其原因也较为明朗:专业型人工智能需要法学专业知识与技术知识的深度融合,且需要在专业话语与技术话语的权力冲突中找到平衡点,故专业型人工智能是需要法学家与人工智能专家共同开发的全新人工智能。显然,现阶段的专业型人工智能并没有实现上述知识的深度融合。

故此,司法领域常规型人工智能发展态势较好,而专业型人工智能效果不理想是现阶段司法人工智能的中国处境。考虑到行文焦点,本文以下所讨论的专业型人工智能主要指介入量刑机制的司法人工智能(以下简称"量刑人工智能")。

(二)困境剖析:法学实证研究的理论供给不足

如上所述,现阶段司法领域专业人工智能无法达到预期效果是我国司法人工智能的现实困境。造成此困境的原因并非人工智能在司法领域存在技术瓶颈,也并非人工智能技术知识与法学专业知识具有原生的不相容性,而是法学实证研究对司法领域专业型人工智能的理论供给不足。晚近以来,我国的法学研究大多是"黑格尔式"思维的规范性研究,实证研究极少,而以大数据分析为根基的量刑人工智能正是以法学实证研究为基础。

第一,量刑人工智能以大数据分析为基础。量刑人工智能以现有量刑生态为基准,以对法官集体智慧的认同为前提,具体表现为对现存已经判决的量刑数据为数据来源进行大数据分析。以现有量刑生态为基准,"意味着司法人员的彼此认同和信赖,因而也意味着司法实践良性的整体延展"[6]。大数据分析更多的是运用统计学原理,也即量化研究;而传统的规范法学研究以定性研究为基本范式,因此,对于习惯于定性研究的法学领域,基于大数据分析的量刑人工智能构建完全是一个陌生的领域。也正因如此,左卫民教授在探索大数据法律研究的科学方式时提出,要"推动研究的团队化与多学科的交叉融合,并致力于培养复合型大数据法学人才"[7]。

第二,量刑人工智能并非人工智能对量刑数据的常规分析。在通用领域,对某一对象的大数据分析已有成熟的技术沉淀,但遗憾的是,已有的技术沉淀并不能直接平移至司法领域,现有量刑辅

助系统的实践困境即为例证。量刑领域的大数据分析首先需要对半结构化的判决书进行数据提取,包括情节、刑罚量等。其次,需要全样本分析,而不能仅以现有裁判文书网的判决书为数据来源。在统计学理论中,以随机抽样为前提的样本统计量通常被认为在一定程度上可以代表总体参数,但值得注意的是,裁判文书网公开的案件量与实际案件量具有一定的差距,且公开的案件并非是在实际案件(总体)中抽样得出,因此以裁判文书网的数据揭示的现象是否能够代表我国的司法现状值得怀疑。“最高法院在2016年之前一共公开了11 080份裁判文书,相较于最高法院2014年与2015年的总体案件审结量,经过计算比例约为46.13%”[8]。再次,需要以量刑理论为基础和指引进行大数据分析。最后,需要利用量刑理论对分析结果进行专业解释。

第三,精细的量刑实证研究是量刑人工智能研发的理论根基。由以上两点可以得知,量刑人工智能的构建需要量刑理论与人工智能理论(底层为数据分析理论抑或统计学原理)深度融合,而深度融合的基础即是具象的量刑实证研究。在以往的法学研究中,也有学者应用定量研究的方法,例如白建军教授就一直耕耘于法学实证研究领域。与以往法学实证研究不同的是,量刑人工智能实现了从“大量数据”到“大数据”的跨越。

综上,现阶段关于人工智能介入量刑机制的研究呈现出抽象有余而具象不足的特征,这就导致量刑实证研究相关理论对量刑人工智能的构建理论供给不足。也正因为如此,才导致现今量刑人工智能在司法实践中“遇冷”。

二、宏观定位:辅助量刑抑或决定量刑

关于人工智能介入量刑机制的宏观定位,“辅助量刑”和“决定量刑”是两种不同的立场,总体而言,“辅助量刑”立场主张在人工智能介入量刑机制之后作为“人”(碳基)的法官仍然是居于主导地位;“决定量刑”的立场则认为人工智能(硅基)应该在量刑过程中居于决定性地位,而具有生物特征的“人”只具有辅助性作用甚至将其排除在量刑程序之外。显然,这两种截然不同的立场仅仅是基于不同的逻辑前提(时空背景)。故此,本部分以现阶段人工智能的发展程度为时空背景,剖析现今人工智能介入量刑机制的定位及其介入的具体方式(参考性抑或规范性)。

(一)逻辑前提:量刑人工智能的时空背景

不可否认,当下学界无论是对人工智能刑事责任能力的讨论,还是对人工智能裁判的讨论都存在一定程度的设想,即以人工智能具有意识或者在可预见的未来具有意识为假设前提。刘艳红教授提出警醒,“人工智能技术热潮的再度兴起,使得人工智能法学研究空前繁盛,但当前研究出现了违反人类智力常识的反智化现象”[9]。笔者认为,现阶段关于人工智能的法学研究要以现今及可预见未来的人工智能技术背景为逻辑前提,而对人工智能发展阶段的认定是法学学者无能为力的,只能求助于人工智能技术专家。故此,本文以当下世界人工智能学者对人工智能的研究现状为逻辑前提作为讨论的基础③。

麻省理工学院物理学终身教授、未来生命研究所创始人马克斯·埃里克·泰格马克(Max Erik Tegmark)将生命发展区分为三个阶段:“生命1.0(生物阶段),靠进化获得硬件和软件;生命2.0(文

③本文并不否定某些领域一定程度的超前研究的价值,例如人工智能有益运动支持者(Members of The Beneficial-AI Movement)所主张的人工智能安全性研究。

化阶段),靠进化获得硬件,但大部分软件是由自己设计的;生命 3.0(科技阶段),自己设计硬件和软件"[10]37。显然,人类处在生命 2.0 阶段,我们可以通过学习不断升级自己的软件,但硬件的升级只能依赖于生物进化。生命 3.0 被人工智能学者认为是人工智能发展至智能爆炸(奇点)时的生命形态④,而当下刑法学者所讨论的具有刑事责任能力以及能够完全智能裁判的人工智能即为生命 3.0 形态。

关于生命 3.0 的讨论主要有两个核心问题:第一,何时到来?第二,出现会有什么后果?围绕这两个问题,存在三个学派:数字乌托邦主义者(Digital Utopians)、技术怀疑主义者(Techno-Skeptics)和人工智能有益运动支持者(Members of The Beneficial-AI Movement)。数字乌托邦主义者认为,"数字生命是宇宙进化自然而然、令人期待的下一步,如果我们让数字智能自由地发展,而不是试着阻止或奴役它们,那么,几乎可以肯定地说,结果一定会是好的"[10]40;技术怀疑主义者认为,"建造超人类水平的通用人工智能(可完成任何认知任务,并且完成得至少和人类一样好的人工智能——引者注)实在太困难了,没有几百年的时间,根本无法实现,因此没必要杞人忧天"[10]39;人工智能有益运动支持者则认为,强人工智能在 21 世纪内会出现,但其后果仍然不确定,因此当下对人工智能的安全性研究尤为重要(见图 1)。

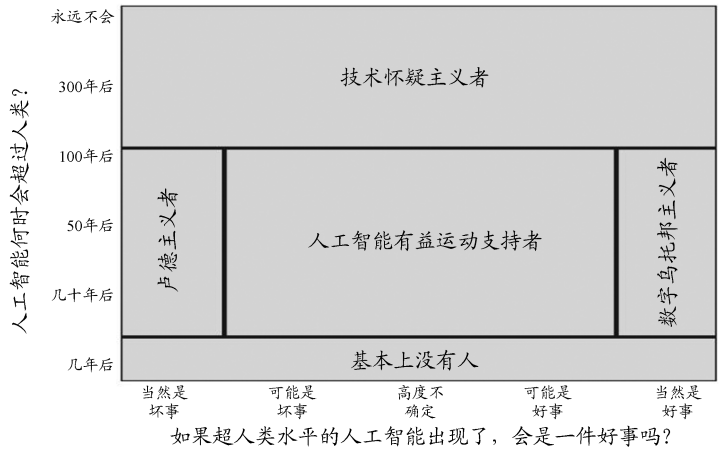

图 1　数字乌托邦主义者、技术怀疑主义者和人工智能有益运动支持者的基本观点[10]42

由以上三个学派的争论可以看出,现今世界最顶尖的人工智能学者对于强人工智能的出现时间各执一词,唯一能够达成的共识是:在当下及以当下存在的人工智能为基础所预见的未来不会出现通用人工智能(强人工智能)。换言之,在当下及可预见的未来,我们仍然会处在生命 2.0 阶段,这正是本文行文的逻辑前提。结合本文焦点,人工智能介入量刑机制现阶段应定位于辅助量刑。当然,为避免陷入"科林格里奇困境(Collingridge' s Dilemma)"⑤,目前对人工智能的安全性控制进行一定程度超前的规范与伦理研究是本文所认同的。

(二)辅助向度:参考性辅助抑或规范性辅助

或由于对强人工智能的忧虑,或由于对当下人工智能技术的清晰认识,抑或由于对"人类中心

④智能爆炸,是指能迅速导致超级智能的迭代式自我改进的过程。

⑤科林格里奇困境,是指一种技术的社会后果不能在技术生命的早期被预料到。然而,当不希望的后果被发现时,技术却常常成为整个经济和社会结构的一部分,以至于对它的控制十分困难(Collingridge D. the Social Control of Technology[M]. Milton Keynes, UK: Open University Press,1980. 11, 11, 16 - 17, 11, 11, 12.)。

主义”的坚守,极少有学者主张人工智能具有决定量刑的地位,大多都认同应将其定位于辅助量刑。但对量刑人工智能的地位认识仅停留在辅助量刑则仍不具有指引价值,甚至会因为其含糊其辞的定位使得司法实践标准不一。量刑人工智能的辅助地位必须进一步精细化,即其是参考性辅助还是规范性辅助,前者是指量刑人工智能给出的“量刑”对法官只有参考价值,后者是指量刑人工智能给出的“量刑”具有规范作用,法官必须采纳或者如不采纳则需要给出合理的理由。笔者认为,规范性辅助才是符合我国量刑规范化改革价值内涵的量刑人工智能定位,而参考性辅助仍然需要最大限度地发挥法官个人的正义与良知,不具有规范属性。值得注意的是,规范性辅助并非是量刑人工智能直接给出法官个案的宣告刑且法官必须适用,而是如本文第三部分关于人工智能介入量刑机制的具象剖析中所详细论述的,量刑人工智能给出的是阶段性量刑,而且给出的是具有“共性”属性的阶段性量刑,如个罪的量刑基准,而具有“个性”属性的量刑则由法官依据裁量权裁量得出,如个案的责任刑、宣告刑等。

第一,规范性辅助符合量刑规范化的价值内涵。量刑规范化改革是近年来我国最重要的司法改革之一,其旨在实现量刑之“规范”,由传统估推式量刑向精细化量刑转变。但遗憾的是,当下从犯罪事实(存在)到具体个案刑罚量(应当)的危险跨越很大程度上仍然依赖于法官的良知与正义。近年来人工智能技术指数爆炸式增长为量刑规范化改革提供了可行的实现路径。量刑人工智能为法官提供某些阶段的量刑(例如量刑基准),使法官能在相对于法定刑较小的幅度内行使自由裁量权,即达到了量刑程序规范化的目的,也实现了量刑实体的规范化。但是,这一切的前提是量刑人工智能给出的量刑具有规范属性,如果其只具有参考属性,则法官所面对的仍然是整个法定刑,量刑过程也无法实现规范化,因为从法定刑选择到最终宣告刑的确定的整个过程仅存在于法官的思维中,无法现实化。

第二,量刑是“刑之量化”基础上的“刑之裁量”。如果没有“刑之量化”,量刑就是无根基、无标准之裁量,其也就成了全部依赖法官正义与良知的无言之知。“刑之裁量”与“刑之量化”并非截然不同的对立立场,而是量刑不同阶段的侧重点。“刑之裁量”强调量刑过程中法官自由裁量权的发挥,“刑之量化”一则强调量刑过程的规范性,二则强调量刑实体的精准化。法官在量刑人工智能给出的“刑之量化”的基础上进行“刑之裁量”,量刑才能科学化、规范化。法官基于参考性辅助所做的量刑并非是以“刑之量化”为基础,而最终仍然仅以自身的正义与良知为基准。另外,规范性辅助与参考性辅助的功能性差异会导致“刑之量化”的标准尺度存在差异,如果将功能定位于参考性辅助,则“刑之量化”的标准将大幅降低。

第三,“量刑均衡是量刑统一化与量刑个别化的辩证统一”[11]。“量刑统一化”即“同案同判”或“类案类判”,例如个罪中,都有统一的量刑基准⑥。“量刑个别化”是指由于每个行为人的人身危险性不同,则基于人身危险性的预防刑也不同⑦。“量刑统一化”要求“类案类判”,需要法官检索区域内甚至是全国范围内同类案件的量刑情况,这对于碳基的人来说无疑是巨大的工作量甚至是不可能实现的,但对于基于大数据分析的人工智能而言则是轻而易举。因此,在量刑人工智能得出的量

⑥由于讨论焦点原因,本文不对量刑基准作更为细致的论述,笔者认为,量刑基准是非裁判性的,是犯罪常态的常态量刑。

⑦有学者也在责任刑意义上使用量刑个别化,此种意义的量刑个别化是基于不可能会有两个犯罪行为是一样的,但通说仍然是在预防刑(人身危险性)意义上使用量刑个别化概念。

刑更精准、涵射的数据更全面时,法官没有理由不以此为依据。例如,对于量刑基准而言,由于同一罪名具有统一的量刑基准,量刑人工智能所得出的量刑基准对于法官量刑具有规范意义而不仅仅是参考意义。

第四,量刑人工智能的规范性辅助地位须上升至立法层面才能强制适用。我国《宪法》和《刑事诉讼法》都对审判权的行使作了专门的规定。虽然量刑人工智能的规范性辅助地位没有对法官的审判权核心进行入侵,但其仍然在一定程度上缩小了法官审判权的裁量空间,因此,如果要将量刑人工智能定位于规范性辅助地位就必须进行立法规定。也正因如此,最高人民法院制定的《量刑指导意见》也仅具有指导意义,并不具有强制执行效力。"《量刑指导意见》确定的各项量化标准也不具有法律效力,主要起指导作用,还需要在实践中进一步检验其科学性、合理性,并不断予以修改、完善"[12]37。

以当下及可预见未来的人工智能技术为背景进行讨论是较为务实、严肃的法学研究。因此,对现阶段的量刑人工智能应定位于规范性辅助量刑。

三、微观解构:人工智能介入量刑机制的具象剖析

在明确人工智能介入量刑机制的体系地位之后,对其进行深入而具体的分析更为必要,如此才能让其宏观定位有立足的根基,才能让依托于人工智能的量刑规范化改革落到实处,而不仅仅是在宏观层面进行抽象的表述。

(一)基本立场:量刑逻辑主导算法逻辑

一直以来,我们都存在一种错误的观念,即量刑人工智能的研发是人工智能专家的事,由他们设计算法进行量刑的大数据分析,并最终得出一个统计值。正是由于这样的错误观念或者至少是在某种程度上被这种观念所影响,进而导致当下的量刑人工智能和司法实践需求存在一定程度的脱节。笔者认为,在人工智能作为"客体"的生命 2.0 阶段,即使其具有深度学习能力,其算法逻辑仍然是由作为碳基的人主导。具体到量刑领域,量刑人工智能的算法逻辑应由量刑逻辑主导。

量刑理论为量刑人工智能的研发提供需求指引。任何人工智能的研发都是以需求为导向的,如果量刑人工智能仅仅是对现存的量刑数据进行抓取,然后得出一些常规的统计值(例如某罪名的平均量刑),则该统计值对法官量刑的价值并不大。故此,量刑理论要深入算法设计的每一步。笔者认为,基于责任主义原则,一个规范的量刑过程分为量刑基准的确立—(不法刑的确立)—责任刑的确立—预防刑的确立—宣告刑的确立⑧。没有量刑理论的指引,量刑人工智能只能得出量刑数据的常规统计数据,无法达到规范性辅助的目的,只具有参考价值。

综上,量刑人工智能的研发是法学专业知识与人工智能专业知识的深度融合,且在法学专业与技术的话语冲突中,应由法学专业话语权主导人工智能技术话语权。因此,法学专家要深入到量刑人工智能研发的各个阶段,而不能抱以拿来主义心态。

(二)作用领域:终局性量刑抑或阶段性量刑

量刑人工智能给出的结论可以区分为终局性量刑和阶段性量刑,前者是指量刑人工智能直接

⑧近年来,关于量刑方法有诸多观点,如《量刑指导意见》中的"量刑起点—基准刑—宣告刑"的量刑步骤。

给出最终量刑,也即宣告刑;后者是指量刑人工智能给出的结论是各个阶段的量刑,如给出个罪的量刑基准、不法刑等。当下关于人工智能辅助量刑的讨论几乎都默认其给出的是终局性量刑,即直接给出宣告刑供法官参考,或者直接给出以不经数据清洗的宣告刑为基础的"大量数据"所得出的统计值,例如给出某省盗窃罪的平均量刑。但量刑人工智能直接给出终局性量刑存在如下缺陷:(1)前文论证了人工智能对量刑的辅助应该是规范性辅助,如果量刑人工智能直接给出终局性结论,则违反了我国审判权的专属性原则,侵犯了法官的裁量权。(2)现阶段的人工智能技术还没有给出终局性量刑的能力。量刑是一个极为复杂的过程,是量刑统一化和量刑个别化的辩证统一,而量刑人工智能所依赖的大数据分析技术只能得出量刑统一化层面的结论;量刑个别化是需要即使具有深度学习能力的人工智能也不具备的正义观、良知以及对社会公序良俗的价值判断,而这些需要社会化的人才能做到。因此,量刑人工智能的作用领域应该是阶段性量刑。

即使量刑人工智能给出的是阶段性量刑,但如果其给出的是所有阶段的量刑,那其与终局性量刑仍然别无二致。如前所述,基于量刑人工智能的特有属性和我国审判权的专属性,量刑人工智能只能介入量刑过程中需要量刑统一化的阶段,而量刑个别化的阶段则由法官基于正义与良知自由裁量得出。量刑人工智能至少能够在量刑基准与不法刑的确立阶段进行规范辅助,而对于责任程度的判定及人身危险性的判定则仍然是法官行使自由裁量权的范围,但量刑人工智能可以给出一定的不具有规范性质的参考值,例如美国的 COMPAS 系统。"'COMPAS'的全称为'Correctional Offender Management Profiling for Alternative Sanctions',可译为'罪犯矫正替代性制裁分析管理系统',是由一家名为 Nortpointe 的公司为法院开发设计的,它可根据对犯罪者的访谈和来自司法部门的信息来评估再犯的风险,旨在帮助法官做出更好的或者至少是以数据为中心的司法决策"[13]。

(三)规范辅助:量刑基准与不法刑的算法逻辑

量刑基准与不法刑的确立是体现量刑统一化的量刑阶段,其都是基于已经存在的不法事实,而客观的不法事实是可量化的;责任刑(由不法刑乘以法官裁量的责任系数)与预防刑的确立则需要价值评价,是人工智能无能为力的。即使一些技术专家认为人工智能也能做到价值评价,但这是无法验证的;虽然法官的价值评价也是无法验证的,但其具有程序的正当性,是共同体基于"共情"赋予其正当权力。由量刑理论主导的量刑基准与不法刑确立的算法逻辑如下:第一,量刑基准。量刑基准是一个较为传统的论题,但遗憾的是至今仍是一个较为混乱的概念,甚至存在与"裸刑""量刑起点"等概念混用的情形。笔者赞同张明楷教授所主张的量刑基准是"犯罪的常态"[14]的刑罚量⑨。因此,笔者认为量刑基准是区域性常态不法的常态量刑,而非容易受极端值影响且无法解决不同刑种之间无法量纲化问题的裸刑均值。由此可以得出,量刑基准是非裁量性结果,其是脱离于(先于)个案存在的,也体现着量刑统一化。基于此量刑逻辑,人工智能计算量刑基准的算法逻辑,首先以限定时间范围内、一定区域内无人身危险性评价、无责任降低评价的所有案件为数据源。将数据范围限定在一定的时间范围内是因为要考虑规范的时效性,如不能将 97 刑法生效前的案件作为当下个案的指引;将数据范围限定在一定区域内是因为量刑具有区域性,一定区域内的量刑只能以本区域内的量刑为基准,如入罪标准在各省份之间具有弹性就是量刑具有区域性的体现;之所以要限定

⑨张明楷教授所书原文为"量刑时,应当按照犯罪的常态确定量刑起点",但又在后文中在量刑起点层面讨论"中线论""分格论""形势论"等学说,故此处的量刑起点即为本文所讨论的量刑基准。

无人身危险性评价、无责任降低评价的案件,是因为量刑基准的事实根据为常态不法,而现阶段的判决书并没有对各个量刑阶段进行严格的区分,而是直接给出宣告刑,因此需要对案件进行数据筛选。其次,对量刑情节进行提取并进行次数统计,得到常态不法,即不法的众数⑩。值得注意的是,传统量刑基准理论都默认一个法定刑幅度内只有一个量刑基准,其实不然,一个法定刑幅度内可能具有多个量刑基准,因为一个法定刑幅度内不法的曲线分布可能存在多峰分布,也即多个众数,多个常态不法。传统量刑基准理论默认一个法定刑幅度内只有一个量刑基准主要是因为法定刑的制定也并非以自下而上的实证分析为基础,且传统量刑理论的研究范式也是自上而下的逻辑推理。最后,对该常态不法的量刑进行次数统计,得到量刑众数(常态量刑),即量刑基准。

第二,不法刑。不法刑是在量刑基准的基础上,对个案中相较于常态不法的不法程度更重或更轻所对应的刑罚量。值得注意的是,传统理论一直将故意和过失作为主观要素,且将"客观"作为不法不可动摇的属性,这就使得故意、过失、目的等要素无法融入不法内涵。但在目的行为论犯罪论体系之后,故意、过失、目的等就作为了主观不法要素。长久以来,"对学习刑法的人而言,所有解决方案,似乎总在客观说与主观说之间纠缠不清"[15],故意、过失、目的等作为一种存在,其仍然可以被理解为是"客观"的。本文主张故意、过失、目的等作为不法要素,而在责任阶层则考察刑事责任能力、违法性认识及期待可能性。不法刑确立的算法逻辑,首先将不法事实划分为两类,其一,可以数量化的不法事实,如盗窃金额;其二,不能数量化的不法事实,如未遂情节。其次,进行不法事实与刑罚量的回归分析,得出其相关关系。最后,通过情节变量与刑罚变量的相关关系得出不法刑。需要说明的是,相关关系并非因果关系,因此需要法官作进一步释明。以上步骤涉及人工智能的自然语言处理(NLP,Natural Language Processing)、深度学习(DL, Deep Learning)及数据分析。

量刑人工智能所给出的量刑基准和不法刑对法官量刑具有规范辅助的效果,法官除可以说明合理理由外必须适用;而在责任刑与预防刑领域,需要法官基于价值评价进行自由裁量。如此,才能实现"刑之量化"基础上的"刑之裁量",才能实现量刑统一性与量刑个别化的辩证统一。

(四)公正保障:算法公开与结论的可解释性

在人工智能辅助量刑的研究中,"算法黑箱(black-box)"与"算法歧视(算法偏见)"一直是学界所忧虑的问题。算法黑箱是指,"在人工智能输入的数据和其输出的答案之间,存在着我们无法洞悉的'隐层',即'黑箱'"[16]。"算法偏见问题引发了广泛关注,不同程度损害公众基本权利、经营者竞争性利益和特定个体的民事权益,亟需规制"[17]。为确保量刑人工智能的正当性,"算法黑箱"与"算法歧视"是必须解决的问题。

第一,算法黑箱。美国威斯康辛州诉卢米斯案是人工智能辅助量刑中"算法黑箱"的典型案例。"2013 年,威斯康辛州以五项罪名指控埃里克·卢米斯(Eric Loomis)与拉克罗斯(La Crosse)驾车射击案有关。卢米斯否认其参与了射击行为,但承认他在当晚晚些时候驾驶了涉案汽车。卢米斯承认了其中两项较轻的罪名——'企图逃避交通官员罪、未经车主同意而驾驶汽车罪'"[18]。法院在对卢米斯量刑时参考了 COMPAS 风险评估工具,判处了卢米斯 6 年有期徒刑和 5 年社区监督(extended supervision)。卢米斯主张法院量刑时依赖 COMPAS 系统侵犯了其正当程序权利,但最终

⑩众数(Mode):统计学概念,是指在统计分布上具有明显集中趋势点的数值,代表数据的一般水平。也是一组数据中出现次数最多的数值,有时众数在一组数中有好几个。

威斯康辛州最高法院维持了原判。COMPAS风险评估工具所使用的算法由于涉及商业秘密不便公开,因此其存在“算法黑箱”。毫无疑问,算法不公开的人工智能有违司法公正,在人工智能对量刑发挥规范性辅助作用时更是如此,即使是参考性刑辅助,其对法官量刑也会有沉锚效应⑪。因此,为保证量刑人工智能介入量刑机制的公正性,其算法必须公开,接受监督。

第二,算法歧视。算法歧视主要有以下类型:(1)“内置性编码造成的歧视”;(2)“数据不完整造成的歧视”[19];(3)“偏见代理的算法歧视”⑫;(4)“特征选择的算法歧视”[20]。内置性编码造成的歧视在算法公开接受监督的情形下不会存在。数据不完整造成的歧视是由于抽样而导致的样本无法代表全体所产生的歧视,这在大数据全样本的情形下也不会产生。而偏见代理的算法歧视与特征选择的算法歧视,是源数据(判决)本身所具有的特征,而并非由算法所导致。

第三,量刑人工智能结论的可解释性。量刑人工智能得出的量刑基准与不法刑对法官量刑具有规范性辅助的效果,但裁判仍然需要体现法官对人工智能的“利用”,且为了防止人工智能漏洞,法官在量刑时需要对量刑人工智能给出的结果进行法理解释,包括量刑人工智能得出该结论所依据的数据库范围、量刑逻辑等。如果法官在对量刑人工智能给出的结论进行解释时,发现存在漏洞,则需填补漏洞后重新运行得出新结论;如果量刑人工智能不存在漏洞,但基于合理理由不能适用量刑人工智能给出的结论,则法官在说明理由后可以不适用,例如由于法律的修改而不适宜再基于先前的判决进行当下的量刑。法官对量刑人工智能结论的“改判”也是对数据的实时更新,社会的变迁也就体现在了量刑数据的迭代中。法官需要对量刑结论释明的另一原因是因为量刑人工智能所做的回归分析只能发现变量之间的相关性,而非因果性,故需要对变量之间的相关关系作进一步解释,才能作为合法结论;尤其在人工智能进行深度学习能够自动计算最佳权重后,法官更需要对结论进行释明。

算法公开与量刑人工智能给出结论的可解释性是人工智能介入量刑机制的公正保障,缺一则会导致司法公信力缺失和裁判的不可预测性。

四、结语

依据迈克斯·泰格马克教授关于生命阶段的区分及现今世界人工智能专家对智能爆炸的预测,我们将会在很长一段时间处在生命2.0阶段,距离强人工智能的出现以当下的人工智能发展现状仍然是无法预期的,故在宏观定位层面,人工智能在量刑领域的地位应定位于“客体”,具有规范性辅助地位。在具象层面,量刑人工智能的研发应以量刑理论主导算法逻辑,并且需要保证算法公开与结论的可接受性。

人工智能技术与量刑理论的深度融合是量刑规范化改革继续深化的重要契机,但不可否认的是,其仍然有诸多困境需要突破。首先,量刑人工智能的构建需要量刑实证研究理论的持续供给,而当下我国量刑实证研究仍然属于小众领域。其次,量刑规范化需要进一步完善量刑说理机制。我国刑事判决书着重于对罪名的定性,而对量刑则并无详尽的说理机制,仅仅在判项中予以公布,这也就导致量刑人工智能的构建需要大量严苛的数据筛选和清理,故司法实践中阶段性量刑说理

⑪沉锚效应:心理学名词,指人们在对某人某事做出判断时,易受第一印象或第一信息支配。

⑫偏见代理的算法歧视,又称关联歧视,是指虽然算法设计者采取客观中立的数据,但数据之间形成组合会产生歧视的后果。

机制的构建极为必要。最后,跨学科的理论研究极其紧迫。量刑理论与人工智能理论的知识割裂是当下量刑人工智能开发的现实瓶颈,只有二者深度融合才能使量刑规范化改革发生质的飞越。

参考文献:

[1]刘艳红.人工智能法学的"时代三问"[J].东方法学,2021(5):32-42.

[2]左卫民.热与冷:中国法律人工智能的再思考[J].环球法律评论,2019(2):53-64.

[3]马长山.数字法学的理论表达[J].中国法学,2022(3):119-144.

[4]钱大军.司法人工智能的中国进程:功能替代与结构强化[J].法学评论,2018(5):138-152.

[5]刘子阳,王芳.铆足马力创新发展深化智能应用:最高法信息中心主任许建峰详析智慧法院建设[N].法制日报,2018-03-13(8).

[6]白建军.基于法官集体经验的量刑预测研究[J].法学研究,2016(6):140-154.

[7]左卫民.迈向大数据法律研究[J].法学研究,2018(4):139-150.

[8]马超,于晓虹,何海波.大数据分析:中国司法裁判文书上网公开报告[J].中国法律评论,2016(4):195-246.

[9]刘艳红.人工智能法学研究的反智化批判[J].东方法学,2019(5):119-126.

[10]迈克斯·泰格马克.生命3.0:人工智能时代人类的进化与重生[M].汪婕舒,译,杭州:浙江教育出版社,2018.

[11]石经海."量刑规范化"解读[J].现代法学,2009(3):104-112.

[12]熊选国.《人民法院量刑指导意见》与"两高三部"《关于规范量刑程序若干问题的意见》理解与适用[M].北京:法律出版社,2010.

[13]朱体正.人工智能辅助裁判的不确定性风险及其防范:美国威斯康星州诉卢米斯案的启示[J].浙江社会科学,2018(6):76-85,157.

[14]张明楷.犯罪常态与量刑起点[J].法学评论,2015(2):1-10.

[15]许玉秀.主观与客观之间:主观理论与客观归责[M].北京:法律出版社,2008:3.

[16]许可.人工智能的算法黑箱与数据正义[N].社会科学报,2018-03-29.

[17]刘友华.算法偏见及其规制路径研究[J].法学杂志,2019(6):55-66.

[18]江溯.自动化决策、刑事司法与算法规制:由卢米斯案引发的思考[J].东方法学,2020(3):76-88.

[19]刘雁鹏.智慧司法中的忧虑:想象、剖析与展望[J].理论与改革,2020(3):168-181.

[20]郑智航,徐昭曦.大数据时代算法歧视的法律规制与司法审查:以美国法律实践为例[J].比较法研究,2019(4):111-122.

Artificial intelligence intervention in sentencing mechanism: dilemma, orientation and deconstruction

ZHEN Hang

(*Law School, Southwest University of Political Science and Law, Chongqing 401120, P. R. China*)

Abstract: Judicial artificial intelligence is divided into conventional artificial intelligence and professional artificial intelligence, the former is to directly transplant artificial intelligence that has developed and matured in the general field to the judicial field without special algorithm updates, the main purpose is to liberate judges from heavy "transactional" work, so they cannot intervene in the core content of trials; The latter is an artificial intelligence specially developed for the judicial field, such as the sentencing assistance system, which intervenes in the substance of the trial, which is the core of judicial artificial intelligence. The current practice

status of judicial artificial intelligence is that conventional artificial intelligence is effective because of its solid foundation, but the development and use of extremely important professional artificial intelligence is not ideal, mainly because the research and development of professional artificial intelligence lacks the theoretical supply of legal research, which is concretely manifested as "too abstract but not enough concrete". The deep reason is that there is no deep integration of legal expertise and artificial intelligence technology knowledge, that is, "those who understand technology do not understand law, and those who understand law do not understand technology". At the macro level, under the space - time background of the unpredictable explosion of intelligence, the life 2.0 stage (cultural stage) or the era of weak artificial intelligence is still the stage in the present and foreseeable future for a long time, and the sentencing artificial intelligence pretending to be a "tool" should still be positioned to assist sentencing rather than determine sentencing, and based on the value connotation of the standardized reform of sentencing. It should be further oriented to normative aid rather than reference aid, the difference between the two is the binding force of the phased sentencing conclusion given by the intelligent system on the judge. At the micro level, the algorithm construction of intelligent sentencing system should be based on the principle of sentencing logic-led algorithm logic, and take phased sentencing with "common" attributes such as sentencing benchmark and illegal punishment as the field of action rather than the final sentencing conclusion (declaration of punishment) outside its ability. In addition, in order to prevent algorithmic black boxes, algorithmic discrimination and correlation replacing causation, it is necessary to make the algorithm of sentencing artificial intelligence open and the interpretation of phased sentencing conclusions.

Key words: artificial intelligence; standardization of sentencing; intelligent justice; sentencing assistance; algorithm black box; algorithmic discrimination

(责任编辑 胡志平)