

论“算法偏见”对刑事正当程序的冲击及应对

张沈锲

摘要：刑事司法中的智能辅助系统运用日益拓展，算法偏见对刑事司法的冲击值得关注。算法偏见植根于算法价值非中立的固有属性之中，不宜与算法黑箱问题简单混同，无法以常见的“增强透明与归责”规制路径一并处理。算法偏见常见于无意识的思维框架、视角盲点之中，在信息不对称之外有着固化、锚定、确认偏见三种形态。在刑事司法中，算法偏见来源于类案分析思路的僵化。建议回归正当法律程序，对算法技术祛魅，保障算法偏见的发现，匹配量刑决策的实际思维过程开展纠偏工作，保障被追诉人提出异议的能力辅助和程序权利，引入类比推理，让算法适用类案检索“统一法律适用”路径，更精确地匹配司法决策的需求。

关键词：算法偏见；算法黑箱；类案分析；正当法律程序；技术祛魅

[中图分类号] TP18; D925.2 [文献标识码] A [文章编号] 2096-6180 (2025) 05-135-24

引言

随着数据技术在刑事司法领域的拓展应用，检察机关纷纷引入数据大模型研判，以大数据赋能办理生态环境犯罪、虚假诉讼罪案件中的线索排查、证据收集、类案监督等工作，运用人工智能辅助刑事司法裁判的前景越来越诱人。人工智能刑事司法运用研究已经从“人工智能应不应当引入司法裁判”逐渐扩展到“人工智能如何应用到司法裁判之中”，与之相伴的“算法黑箱”与“算法偏见”问题在刑事诉讼中初步暴露。在算法偏见问题以往的讨论中，学界研究更多集中于新闻服务推送、劳动者平等待遇等民商事法律领域，但是现阶段，刑事司法领域的算法偏见隐忧也越来越浮出水面。结合各国已有的刑事司法人工智能运用，如美国的罪犯矫正替代性制裁分析管理系统(Correctional Offender Management Profiling for Alternative Sanctions, COMPAS, 下称罪犯矫正分析系统)、英国的犯罪评估系统(Offencing Assessment

【作者简介】张沈锲，法学博士，江西财经大学法学院讲师，江西财经大学纪检监察研究中心研究员。

【基金项目】江西省高校人文社会科学重点研究基地 2023 年度研究项目“我省推行涉案企业合规的程序分层问题研究”（项目批准号：JD23098）。

（致谢：2023 年武汉大学第六届网络法治高端论坛提供的宝贵交流机会，江西财经大学法学院章麟同学对本文校对工作亦有贡献。）

System, OASys)⁽¹⁾, 外国系统的算法偏见问题已经存在。以一项对美国罪犯矫正分析系统的研究为例, 特定人群的再犯罪风险预测准确率偏低, 将罪犯矫正分析系统预测为低再犯风险的区间内(再犯风险得分八分以下的区间内)的数据, “与西班牙裔美国人实际的暴力犯罪再犯数据对比, 罪犯矫正分析系统的准确率只有 14%, 即十次预测中至少有八次是错误的”。⁽²⁾ 该研究认为, 美国的罪犯矫正分析系统对西班牙裔美国人存有算法偏见。“随着美国国内对减少累犯以及循证(evidence-base)的日益关注, 此类工具的使用现已扩展到量刑领域。”⁽³⁾ 2016 年的美国合众国诉卢米斯案件(下称卢米斯案)⁽⁴⁾ 也是典型的算法偏见问题, 后文将详细讨论。为聚焦主题, 下文中所称的算法偏见是指在刑事司法智能辅助系统中, 算法受到基础数据、算法设计等因素影响, 固化、增强甚至重新生成对特定人不公正的倾向。因为以计算机技术语言的面目出现, 该类偏见往往难以为被追诉人所理解与发现, 也不易为刑事司法裁判人员察觉与纠正。算法偏见也将影响司法责任的落实, “对于官员来说, 形式上的中立性就足以成为卸责的理由。还有的公共部门会用其他替代性指标‘虚饰’歧视性指标, 以规避问责”。⁽⁵⁾

从我国上海刑事案件智能辅助办案系统(下称 206 系统)问世之日起⁽⁶⁾, 现阶段越来越多的检察监督大数据模型被研发出来, 伴随大数据运用而来的“算法偏见”也越来越需要被认真对待。自 2021 年最高人民检察院推动逮捕羁押“社会危险性评估规范化”的专项试点以来, 量化评估指标体系日益精细化, 有望转化用于算法工具, 辅助羁押必要性审查。不同于民商事领域, 刑事司法领域中人工智能运用一旦拓展, 则可能体制性推行、大规模运用于社会危险性评估、认罪认罚从宽量刑建议提出等工作之中, 既被赋予统一司法裁量的期待, 也潜藏了算法

(1) Andrew Ashworth, *Sentencing and Criminal Justice* (6th ed.), Cambridge University Press, 2015, p. 91.

(2) 该项调研测试原理是将 COMPAS 的预测与实际两年内再犯情况进行对比, 判断 COMPAS 所预测的再犯风险多大程度上与实际的犯罪情况相一致, 该测试运用了 2013—2014 年期间佛罗里达州布劳沃德郡(Broward County, Florida)的犯罪统计数据, 涵盖两年内一般犯罪再犯数据库 6 172 例和暴力犯罪再犯数据库 4 020 例, 布劳沃德郡是美国人口数前二十的郡之一, 2008 年开始布劳沃德郡警察局开始运用 COMPAS 系统。COMPAS 算法生成再犯风险得分按 1—10 计算, 1—4 分是低风险, 5—7 分是中风险, 8—10 分是高风险。Melissa Hamilton, *The Biased Algorithm: Evidence of Disparate Impact on Hispanics*, 56 American Criminal Law Review 1553, 1575 (2019).

(3) 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》, 载《东方法学》2020 年第 3 期, 第 77 页。

(4) 卢米斯(Eric Loomis)被威斯康辛州检察官指控与拉克罗斯(La Crosse)有关驾驶汽车射击案件有关的 5 项指控, 卢米斯否认射击, 但承认“企图逃避警察、未经车主同意开车”有关的两项指控。威斯康辛州的惩教署官员按程序提供审前量刑调查报告时运用了大数据的 COMPAS 工具, 该软件评估方法不公开, 从访谈、犯罪记录中获取信息, 评定卢米斯的累犯风险数值。初审法院采纳判处 6 年监禁和 5 年缓刑(extended supervision)。卢米斯提出反对意见: (1) 侵犯其根据精准信息被量刑的权利, 因为 COMPAS 工具提供的仅是与特定群体的统计数据对比, 而且评估过程不对其披露; (2) 侵犯其基于个案事实获得裁判结果的权利; (3) 侵犯其正当程序权利, 因为该系统将性别问题纳入风险评估范围。该案最终在威斯康辛州最高法院维持原判, 理由经梳理包括: (1) 性别作为考虑因素, 仅基于精确性考虑而非歧视; (2) 该工具的基础数据都是公开裁判结果, 如质疑其精确性, 须卢米斯提供证明; (3) 主笔判决书的 Ann Walsh Bradley 法官也承认该数据并非个性化分析, 而是“与被告相似群体的累犯风险聚类数据(aggregate data)而来, 但同时强调其决策辅助性, 认为经过了初审法官审查、并非唯一依据, 具有合理性”。上述内容综合了案件判决书原文、分析文章。State v. Loomis 881 N.W. 2d 749 (Wis. 2016), <https://cases.justia.com/wisconsin/supreme-court/2016-2015ap000157-cr.pdf?ts=1468415026>.

(5) 高童非:《论自动化行政中的“算法卸责”》, 载《电子政务》2023 年第 2 期, 第 127 页。

(6) 2017 年上海市高级人民法院牵头研发了“上海刑事案件智能辅助办案系统”。参见熊秋红:《人工智能在刑事证明中的应用》, 载《当代法学》2020 年第 3 期, 第 80 页。

偏见的隐忧。刑事司法牵涉对被追诉人人身自由的限制，也深刻嵌入被视为社会公正基石的刑事司法程序之中。鉴于算法偏见问题的重要性，笔者在此予以专题讨论。不同于以往民商事领域的算法偏见研究，笔者聚焦刑事程序法领域，不停留于算法偏见混同于算法黑箱的笼统讨论，分析算法偏见有别于算法黑箱的生成原理，也认为刑事司法“人”的自由裁量不必无条件地顺从于大数据技术的“机器理性”，“人”的因素应当成为最终依靠，矫正算法偏见、维持个案公正。依循上述线索，笔者按照算法偏见类型的划分，考虑算法偏见的形成节点、产生原因，以技术祛魅发现刑事司法中的算法偏见，以类比推理引导裁判思路的回归，以数据权利救济挖掘算法权力制衡的程序潜力。

一、“算法黑箱”之外的算法偏见

有学者认为，找不到价值中立的平等算法，寻找这样一种算法是不可能完成的任务。⁽⁷⁾司法决策要求的公正无偏倚，与算法价值预设或偏向构成紧张关系，算法偏见几乎与生俱来。与算法黑箱的有意利用信息优势不同，算法偏见根源并非信息不对称，也非利用信息优势的故意心态，而是无意识的思维结构缺陷，或者观察视角的盲区问题。有学者分析认为，嵌入人工智能的歧视，既有内隐偏见驱动的无意识歧视，也有结构性不平等驱动的系统性歧视。⁽⁸⁾由此带来的算法偏见，不能单纯依赖“信息透明”予以消除，就其偏见所处位置而言，处于“算法黑箱”之外，亟须更具针对性的分析。本部分结合认知理论，分析算法偏见在三个方面的影响，包括巩固既有偏见、锚定复杂案件的偏见，以及为偏见提供技术与程序验证的效果。

（一）固化偏见

在涉及羁押必要性审查的社会危险性评估中，“很难完全控制住其他自变量对羁押率变化的影响，如逮捕理念、逮捕程序、考核指标等因素很难完全保持不变”。⁽⁹⁾但是，实际的逮捕决定可能是出于某时某地特殊的刑事政策，如果不加区分地使用这些数据开发模型、训练算法，则可能将基础数据中的偏见带入算法之中。例如，有学者研究发现，“目前较为成熟的类案裁判系统，其适用的范围仍然较为狭窄……即纳入到样本库的基本上都是有罪判决”。⁽¹⁰⁾以有罪判决为主训练的算法，算法偏见可能导致主要得出有罪判决的结论，进而影响大数据工具后续的预测结果。“作为一种决策机制，（算法）它只能基于已设定的指令作出程序性判断，而无法掺杂情感、伦理、道德等价值判断。”⁽¹¹⁾但是，算法本身包含了以往裁判数据的偏见，例如美国司法中黑人的逮捕率数据本身就偏高，“可能因为历史上的种族主义、警察实践中的不公或者刑事司法系统中的不平等问

(7) 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，第1615-1618页。

(8) 李成：《人工智能歧视的法律治理》，载《中国法学》2021年第2期，第129-131页。

(9) 高通：《逮捕社会危险性量化评估研究——以自动化决策与算法规制为视角》，载《北方法学》2021年第6期，第138页。

(10) 孙海波：《反思智能化裁判的可能及限度》，载《国家检察官学院学报》2020年第5期，第92页。

(11) 张玉洁：《智能量刑算法的司法适用：逻辑、难题与程序法回应》，载《东方法学》2021年第3期，第199页。

题而被逮捕和羁押”。^{〔12〕}如果原有裁判数据以完整案例形式展现,可能容易识别其中包含的偏见,但是如果以算法运用后的风险概率形式展现在量刑辅助工具之中,则新案件的裁判者更加难以识别偏见。“算法只是简单地作为‘传送带(transmission belt)’将以往歧视性司法偏见复制到后续的保释实务之中。”^{〔13〕}原有偏见变得难以识别,算法起到了固化偏见的作用。

在卢米斯案中,即使运用算法的罪犯矫正分析再精细,也只是统计学意义上的精细,并非真正的个案裁判。美国司法中种族偏见由来已久,罪犯矫正分析则固化了对非洲裔、有色人种美国人的量刑偏见,这是学者提及美国司法偏见时常用的例子。虽然统计学的结论看似客观、理性,实则并不能排除偏见,“算法不仅没有让社会从偏见之中解脱出来,没有避免对法官决策程序的污染,反而延展了社会中已经存在的偏见(entrenched cultural biases that preexist in the society)”。^{〔14〕}在机器学习过程中,算法携带了原有司法裁决的偏见。

即使没有算法工具,司法裁量中的一定程度差异可能也是客观存在的,这是法官自由裁量导致的,在案件事实、被追诉人情况完全一致的情况下,仅仅是“受谁审判”,也可能带来案件裁判中的明显量刑差异。根据学者统计,故意伤害罪中法官量刑差异能达到11.26%,“在不同法官的量刑差异可以达到10%的情况下,一个宣告刑在3年有期徒刑上下浮动的被告人,其可能由于案件被分配给不同的法官审理,而面临4个月左右的刑期浮动。从被告人的视角看,这4个月的刑期浮动与其犯罪事实和人身危险性无关,而仅仅与案件由谁审理有关”。^{〔15〕}算法偏见可能固化、保留甚至重现这一司法裁判者“人”的因素带来的量刑偏差。

(二) 锚定偏见

德国社会学家马克斯·韦伯曾批判过分职业化、官僚化的司法体制,代表普通人说出了对机械司法的反感,即法官仅起到一部加工机器的作用,这种法官只知道将当事人的诉讼要求和诉讼费一起塞入机器,然后根据从法典中推演出的理由进行诊断。^{〔16〕}在工业时代,“法官像机器一样裁判”被作为时代隐喻,有力讽刺着司法机械主义;进入数字时代,这一命题回转过来成为另一时代隐喻——机器也会像法官一样裁判,从一端输入事实,从另一端输出判决。司法智能辅助系统正在逐步“替代”法官的部分裁量功能。在司法领域,尤其是裁决他人生命、自由的刑事司法领域之中,算法是否会喧宾夺主?虽然算法定位为“辅助”作用,但司法决策过程中“非人性”的算法越来越重要,甚至演化为结构性偏见,已经成为需要严肃对待的问题。

〔12〕 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》,载《东方法学》2020年第3期,第80页。

〔13〕 原文为:“Insofar as those factors are themselves reflective of policy decisions of doubtful fairness, the algorithm will simply serve as a transmission belt, replicating discriminatory practices at the subsequent bail stage.” See Vincent Chiao, *Fairness, Accountability and Transparency: Notes on Algorithmic Decision-Making in Criminal Justice*, 15 International Journal of Law in Context 126, 127 (2019).

〔14〕 原文为:“Far from liberating the society from the biases and prejudices that might pollute judges’ decision-making process, these tools can intensify, while simultaneously concealing, entrenched cultural biases that preexist in the society.” See Kia Rahanama, *Science and Ethics of Algorithms in the Courtroom*, 1 University of Illinois Journal of Law, Technology & Policy 169, 169 (2019).

〔15〕 吴雨豪:《量刑自由裁量权的边界:集体经验、个体决策与偏差识别》,载《法学研究》2021年第6期,第120页。

〔16〕 [德] 马克斯·韦伯:《论经济与社会中的法律》,张乃根译,中国大百科全书出版社1998年版,第355页。

大数据技术运用于刑事司法,存在着另一重隐忧,预测风险结果只要被数值化,则无论数值大小,裁判者都倾向于认为风险存在。美国一项调查称,“一个可能的解释是,法官、陪审员可能‘有点数字盲(have issues with innumeracy)’”,尤其是涉及统计数值时”。⁽¹⁷⁾智能辅助系统依据统计学的规律给出建议的“预先决策”,以“数字”“比例”为推理加工工具,完全不同于人作出的“文字叙述性”的价值权衡与推理选择。算法起点是数据库,得出的有效内容只是“预先决策”数值,造成了裁判者的理解困难,为算法偏见留下生存空间。如果陪审团认为实质同一的再犯率做等值变换后(10次中5次再犯,50%再犯率),百分比更直观,则被认为再犯率更高,即第一组以数字的频次“10次中有5次”表示再犯风险,第二组将上述频率做等值变换,标识为“50%的再犯率”,在陪审团看来第一组具有更高的再犯风险。⁽¹⁸⁾如非有实证研究支持,这一结论简直是对陪审团司法准确性的嘲弄。无独有偶,即便同样是以百分比表达,因为描述的角度不同,人身危险性感受也不一样。在另一试验中第一组再犯可能性为26%,第二组不再犯率为74%,则第一组罪犯被认为更危险。⁽¹⁹⁾数字化的表达往往具有科学、精确的外观,只要给出具体指向,即指向某个案件中的再犯率,则即使证明力非常弱,法官也仍倾向于跟随这一指向而作出裁判。对此,行为心理学上称之为“锚定效应”(anchoring effect)——即便一份独特、可理解证据证明力非常弱,人们也会因为其容易理解、容易记忆,而在后续裁判中不自觉地用到这份证据,实际是以数值表达作为具象化工具(actuarial tool),发挥锚定效应影响内心确信的作用。如果回归理性视角,则会认为数值完全无益于把握问题本质,在决定是否要监禁被追诉人的过程中并不具有任何参考价值,百分比表示的危险性并不比简单的高中低的危险性归类更具有决策参考意义,却被执着地采信着。⁽²⁰⁾由此产生的偏见被称为锚定偏见(anchoring bias)。

锚定偏见是指从一个固定点出发,在面对各种不确定性的情况下,越来越依赖这一锚定点,作出最终的判断。⁽²¹⁾锚定点可以用于解释“量化”带来的问题,量化后的“精算结果”(actuarial

(17) Melissa Hamilton, *Adventure in Risk: Predicting Violent and Sexual Recidivism in Sentencing Law*, 47 Arizona State Law Journal 1, 58 (2015).

(18) Paul Slovic, John Monahan & Donald G. MacGregor, *Violence Risk Assessment and Risk Communication: The Effects of Using Actual Cases, Providing Instruction, and Employing Probability Versus Frequency Formats*, 24 Law and Human Behavior 271, 295 (2000).

(19) Nicholas Scurich & Richard S. John, *The Effect of Framing Actuarial Risk Probabilities on Involuntary Civil Commitment Decisions*, 35 Law and Human Behavior 83, 88(2011).

(20) 该理论来源悠久,有大量研究提及了锚定效应对法官审理偏见的分析,为此,有学者专门作了社会心理学的对比实验,通过控制法官知识程度,以“数字化”的假定,模拟法官对过去合理裁判标准的知识,并以此开展了三项对比研究。所谓“数字化”,是指对过去的合理裁判经验,经过统计与量化分析,得出一项标准经验。法官在裁判新案件中,很大程度上会模仿这一标准经验,模仿程度则取决于其对该标准的知悉程度。锚定效应影响裁判过程,法官对案情掌握得越少,法官越可能模仿锚定效应的裁判模式。即使法官知道锚定点是人为设置的,锚定效应也仍然会起作用。Thomas Mussweiler & Fritz Strack, *Numeric Judgments Under Uncertainty: The Role of Knowledge in Anchoring*, 36 Journal of Experimental Social Psychology 495, 515 (2000).

(21) 锚定偏见发生于任何决策过程中,如果决策者对某一信息看得过重,该信息即是锚定点,通常锚定是起点,而且能够维持全程的影响力直至最终作出决定,尤其是在面临不确定事实的时候。锚定偏见是一种启发式的认知,如果面临复杂疑难决策时,依赖一个锚定点看起来是心理上的捷径,但是也可能会无意识地产生重大错误。Mellisa Hamilton, *Adventure in Risk: Predicting Violent and Sexual Recidivism in Sentencing Law*, 47 Arizona State Law Journal 1, 22(2015).

prediction) 无论高低, 都可能会让法官变得更加保守, 因为心理上产生了锚定点, 法官逐渐趋向依赖这一锚定点, 智能辅助系统也就反客为主, 以锚定效应操控着法官形成内心确信的过程。在面临大量不确定事实的情况下, 法官如果采纳辅助决策结论就多了一重支持证据, 法官如果不采纳则承担了比以往更高的心证责任, 从而难以排斥代表着类案经验的“辅助决策结论”, 实际上损害了保障个案公正所依赖的自由裁量。

(三) 确认偏见

确认偏见 (confirmation bias), 是指“如果一项风险预测工具给出预测结果, 无论结果高低, 法官和陪审员只要事前已认为存在高风险, 那么法官和陪审团都可能高估预测结果的意义, 以支持自己的结论”。⁽²²⁾ 这是一种不自主选择与解释证据的趋势, 即高估、确认过去认可的意见和假设, 贬低、否认过去反对的意见和假设, 这实际是一种认知错位。当出现相反证据时, 人们倾向于相信支持已有决定的证据, 压制反对已有决定的有关征兆或证据, 更愿意相信以前的经验在现在的案件中重现, 而懒于自我怀疑, 并且寻找各类证据作证确认这一现象的存在。例如, 有表述称, 处于离异的婚姻状态显著增加多种再犯类型发生的可能性, 以农民工为代表的群体对社会秩序构成了巨大的威胁, 是新增犯罪的主要来源。⁽²³⁾ 上述犯罪统计结果有着特定社会经济条件下的不可避免性, 即有些人因为客观因素处于经济贫困、社会地位低下的状态, 但是不能简单地以统计上的关联关系推断个案的因果关系, 即贫困人口不必然具有更高的犯罪可能性。虽然犯罪者作为一个群体, 多数处于贫困且社会地位低下的状态, 但是不能认定只要贫困且社会地位低下就必然犯罪, 也不能因此在个案中机械适用更长的审前羁押期限。“实际上高风险只是代表一种可能性, 并不意味着凡是符合低学历、农村户口或无工作情形的犯罪嫌疑人、被告人就会实施具有社会危险性的行为。”⁽²⁴⁾ 这类偏见往往更加隐蔽, 难以发现和纠正。

算法偏见是开发主体在无意识状态下、基于认识盲区而形成的, 很难由开发主体与技术人员自行发现和纠正。美国的罪犯矫正分析系统同样属于智能辅助系统, 被认为有着高估再犯风险的趋势。因为对开发者而言, 高估风险比低估风险更安全, 因为更谨慎地防范犯罪对社会而言总没有坏处。“如果风险评估算法高估了罪犯的累犯风险而要求对其采取措施, 这种错误通常很难被发现, 即使被发现也不会给公司造成社会丑闻; 但如果该算法低估了罪犯的累犯风险而建议将其释放, 一旦罪犯再次实施了暴力犯罪, 开发者可能会面临强烈的批评。”⁽²⁵⁾ 长此以往, 在开发者利益驱动下, 罪犯矫正分析系统会越来越趋向于高估再犯风险, 与之类似, 各类算法越来越趋向于

(22) 原文为: Thus, judges and jurors may overvalue a risk tool prediction that confirms their preexisting inclination toward assuming high risk. See Mellisa Hamilton, *Adventure in Risk: Predicting Violent and Sexual Recidivism in Sentencing Law*, 47 Arizona State Law Journal 1, 60 (2015).

(23) 汪晓翔、刘仁文:《不同再犯类型的差异化风险因素研究》, 载《中国人民公安大学学报(社会科学版)》2020年第5期, 第8页。

(24) 高通:《逮捕社会危险性量化评估研究——以自动化决策与算法规制为视角》, 载《北方法学》2021年第6期, 第138页。

(25) 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》, 载《东方法学》2020年第3期, 第79页。

严惩犯罪、轻视权利保障，如果这类算法逐步推广运用于刑事司法程序，算法价值选择的极端化倾向将不可避免地不断积累，是否对刑事诉讼控辩平衡造成结构性影响，值得进一步关注。

“算法偏见”并非特例，而是“歧视数字化”下的普遍趋势。在数据技术及网络运用过程中，歧视无处不在，在个体与社会数据处理的各个环节，都存在着以技术的外观嵌入偏见的可能性——“人工智能的发展引发人类认知的第四次革命，但人工智能中存在着算法偏差，算法偏差可能引起算法歧视”。⁽²⁶⁾ 算法歧视广泛存在，“责任规则、信息约束和平权行动”存在“归责不能、约束失效”的问题。算法偏见与算法黑箱没有必然联系，大数据技术的核心在于算法，“大数据不仅仅是指数据体量、容量和多样性，也包含了适用软件抽取新的、预测性知识，帮助个体或群体决策之目的”。⁽²⁷⁾ 以算法处理司法裁判的数据库，用于评估社会危害性、人身危险性、再犯可能性等司法决策指标，为个案裁判提供参考。在大数据技术运用过程中，算法具有举足轻重的地位。就大数据的含义而言，数据（data）本身只是一种无意义的数字和字符，只有特定的数据组织与解读方式才赋予了数据特定含义，算法（algorithm）即是大数据的数据组织与解读工具。“数据化是其本质特征，统一化是其派生特征。”⁽²⁸⁾ 经过算法的加工提炼，基础数据转变为有决策参考意义的分析结果。司法智能辅助系统运用关键仍在于保障公正，即便司法决策长期传统是强调法官“自由心证”，人工智能的“机器理性”也仍在不断影响法官的自由心证，侵蚀法官的自由裁量空间。为此，在个案司法裁判中，法官的自由裁量须警惕不得过分让位于机器理性。有学者认为，“算法在接管人类让渡的决策权”。⁽²⁹⁾ 司法智能辅助系统本无法完全量度社会情感等非指标化的因素，以数值化简化司法裁量中的价值衡量过程，则不可避免地产生价值衡量偏颇，引入算法偏见，价值衡量本身就是统计与数值化语言所无法处理的，也是无法通过算法的自我推演发现、矫正的。在这个意义上，大数据技术的运用前景，很大程度取决于司法机关能否恰当地发现处理算法偏见，为消除偏见，首先需要追溯偏见来源与成因，进而通过数据技术、统计规律之外的司法制度本来因素对症下药，解决算法偏见问题。

二、刑事司法中算法偏见的成因

有学者指出，“主观性因素普遍存在时，统计分析则无法将这些主观性因素从算法中去除”。⁽³⁰⁾ 算法的“统计学逻辑”之下，以数字技术思维理解司法思维，可能带来统计学对个案特殊情况的理解盲区，单纯依赖数字技术思维永远无法贴近人性，无法为维护个案公正而作出价值衡量、道德判断，更无法预测某一案件裁判对被裁判个人的规范效果。严格地讲，算法偏见是一个复合概

〔26〕 闫坤如：《人工智能的算法偏差及其规避》，载《江海学刊》2020年第5期，第141页。

〔27〕 European Commission for the Efficiency of Justice, *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment*, Council of Europe Portal, <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>.

〔28〕 刘品新、陈丽：《数据化的统一证据标准》，载《国家检察官学院学报》2019年第2期，第129页。

〔29〕 张凌寒：《算法规制的迭代与革新》，载《法学论坛》2019年第2期，第21页。

〔30〕 高通：《逮捕社会危险性量化评估研究——以自动化决策与算法规制为视角》，载《北方法学》2021年第6期，第139页。

念,是指算法设计、实施全过程中对被适用对象各种程度、不同形式的偏见。算法偏见既包括设计者的偏见与认识盲区,也包括基础数据中的偏见,还包括机器学习运用、训练纠错测试过程中的偏见。换言之,算法偏见包括在设计、利用基础数据学习、算法实施结果解释等各环节之中,基于固化错误认识、模糊责任主体、技术与司法语言转换中的可解释性障碍等问题造成的各类偏见。

(一) 算法偏见发现的障碍——固化司法认知缺陷

刑事司法中的算法运用分为三个环节,即证据数据化、数据整合、智能的深化与输出。⁽³¹⁾ 算法偏见也与之对应形成于算法运用的各个环节。“算法偏见萌芽于数据收集步骤,成熟于模型完善步骤,强化于模型应用阶段。”⁽³²⁾ 智能辅助系统服务于刑事司法,与刑事司法的“自由心证”传统存在一定的冲突。因为算法的技术门槛高,司法人员无法直接理解技术语言,仅参考算法决策的产出结果。算法本身则仍是“算法黑箱”、无法被解释与分析,但是司法权的最终判断权属性决定了司法领域不能有无被解释、被分析的判断内容。“尽管学界并不认可算法能够自我进化(self-optimize),但是算法本身确实以超出人类理解能力的方式自我极化。”⁽³³⁾ 算法的自我极化,就是任由算法黑箱演化、迭代,以“功能替代”的方式通过计算机的语言,解决人类裁判者所需要处理的问题。但是由于机器语言与人类语言的非同构性,“电脑编程要求‘具有人类语言沟通所不具备的特别精确、正式、完整程度’,这就导致编程语言对大多数普通人是无法理解的”。⁽³⁴⁾ 即便算法产出结果日趋实质公正,也无法真正回归到人类的裁判思维之中与之对话。

算法形成的过程,即包括了基础数据的训练。刑事司法体系之所以引入大数据技术,是为追求裁判标准的统一适用,大数据技术中算法的生成与训练,都依赖于过去庞大的案例库基础。上海206系统吸纳了大量基础数据作为支撑,仅截至2018年12月底,“公安机关累计录入案件2.48万余件;检察院批准逮捕8811件;检察院审查起诉7442件;法院收案4812件;法院审结3438件”。⁽³⁵⁾ 算法通过统一化、规范化的方式,寻找司法数据的统计学规律。量刑辅助决策工具看似更具结构化、清晰展示其过程科学性的特点,但是这实际上将案件作了抽象、排除了个案特殊性的“人”的考量,算法运用到各国原本只有“人”才能开展的司法决断领域。这一抽象的方式是提炼不同的指标体系,广州南沙区逮捕社会危险性量化评估模型围绕人身危险性、社会危害性、

(31) 栗峥:《人工智能与事实认定》,载《法学研究》2020年第1期,第119-132页。

(32) 刘友华:《算法偏见及其规制路径研究》,载《法学杂志》2019年第6期,第55页。

(33) Jenna Burrell & Marion Fourcade, *The Society of Algorithms*, 47 Annual Review of Sociology 213, 221 (2021).

(34) 原文为: Writing for computational devices demands “a special exactness, formality, and completeness that communication via human languages does not,” and this renders code language “inaccessible to the majority of the population”. See Kia Rahnama, *Science and Ethics of Algorithms in the Courtroom*, 1 University of Illinois Journal of Law, Technology & Policy 174, 184 (2019).

(35) 余东明:《我国首次应用人工智能辅助技术开庭审案》,载人民网2019年1月24日, <http://legal.people.com.cn/n1/2019/0124/c42510-30588997.html>。

诉讼可控性等形成了 43 个指标。^[36]但研究已经指出特定事项是偏离归类量化模式之外的,少数群体价值观同刑法价值观相冲突的案件、风俗习惯同刑事要件相吻合的案件、经济水平与刑事处罚显失公平的案件^[37],都需要“人”的纠偏与矫正,不能任由“机器”简单归类量化下的偏见蔓延。司法体系以大数据的庞大“计算能力”为基础,进行对有限指标的选择,开展精算选择,在选定指标之后算法运用更加透明、连续一致、有逻辑性,但是指标拣选可能出现遗漏,还存在不宜或者无法通过量化形式展现的考虑因素,都会影响算法是否周全无偏颇。过去案例的基础数据,可能包含了既往裁判者的偏见,大数据技术利用基础数据生成具有预测能力算法,已经将过去裁判中的裁判者偏见固化到算法之中。在机器学习的环节,大数据技术的工作原理仍然依赖“自动决策系统”(automatic decision system),即通过预先设定的代码程序,自动得出参考建议。“自动化”既是算法的便捷所在,也意味着算法有脱离控制的可能。算法通过迭代的方式自我进化,实际也是上述“算法黑箱”不断形成的过程,由于不可解释与分析,算法的实际迭代脱离了人的控制,带来技术层面的新偏见。在量刑辅助决策系统中,“辅助”的定位有赖于维护裁判者与智能系统之间的“主从”关系,一旦算法脱离设计者、使用者控制,则算法基于系统性的特点,反而成为一项压制性的力量。尽管人工智能被定位为处于“辅助”工具地位,实际上却束缚甚至妨碍法官在新型案件中的价值衡量,难以保障自由心证。算法结论便捷易得、大量数据支撑、条分缕析的层次结构,都成了阻碍法官摆脱其干扰的最大妨碍。如何发现、纠正算法偏见,始终是算法决策运用到刑事司法中不可回避的难题。

有学者将算法歧视分为“初始算法歧视、记忆算法歧视和神经算法歧视”^[38],本质上是技术运用过程中产生的,则其分解与消除也需要嵌入各个具体的设计运用环节之中。算法偏见的分析与规制,需要探析算法偏见形成的原因。目前,算法偏见的研究多以现象描述的形式展现,较少涉及具体的原因分析。算法偏见形成与算法适用的过程有关,是算法设计与实施过程产生的问题。为此,有必要首先回顾,司法体系根据何种原理引入算法。就设计原理而言,司法机关引入算法,是为了寻找既往裁判中的统计学规律,从实践经验理性总结的角度,为个案裁判的正当选择提供参考。“整体的方向是由所有个体的合力方向牵引,量刑集体经验的建构建立在对大量具体量刑决策进行综合与归纳的基础之上。……在大数据的语境下,法官的量刑实践必然存在着向集体经验中央聚集的趋势。”^[39]上述运用假定了司法裁量遵循应当类似“正态分布”的统计学规律,但是该假定在多大程度上、多大范围内具有有效性并未得到充足的试验验证,何况,能够摆在法官面前、控辩争议的问题,往往具有个案特殊性,无法简单抹除个案差别,能否盲目以符合统计规律

[36] 钟亚雅、招阳:《广东广州南沙区:探索建立社会危险性量化评估机制》,载最高人民法院官网 2020 年 10 月 10 日, https://www.spp.gov.cn/spp/zdggz/202010/t20201010_481455.shtml。

[37] 张玉洁:《智能量刑算法的司法适用:逻辑、难题与程序法回应》,载《东方法学》2021 年第 3 期,第 198 页。

[38] 杨玉晓:《人工智能算法歧视刑法规制路径研究》,载《法律适用》2023 年第 4 期,第 86 页。

[39] 吴雨豪:《量刑自由裁量权的边界:集体经验、个体决策与偏差识别》,载《法学研究》2021 年第 6 期,第 113 页。

上的“正态分布”作为算法开发假定，值得商榷。

即使用于量刑的智能辅助系统是全新的，而训练用数据库也是原有的裁判数据，这就会导致旧案件中的偏见，叠加进入算法设计的基础案件数据库之中，转化为系统性偏差。因为原有裁判数据的各类偏差，如果只是以单个案件裁判的形式出现，是容易被识别并纠正的，但是如果以“案例数据库”的集合形式出现，则经过算法加工，多个案例作为基础数据变换样式成为过往裁判的“类案”经验，原有偏见不会自然消失，却在过往一般裁判经验的加持下，摆在裁判者面前要求其参考，“（个案裁判存在）地区性偏差、人文性偏差、数量性偏差等。为此，这类误差是以类案形式发挥普遍影响的，而不只是影响个案”。^{〔40〕}这导致算法外观上公平，但所依据的基础数据有偏差，算法运用效果中也自然保留了偏见发挥作用的余地。美国保释基础数据之所以存在偏见，一项重要原因是其保释官常有在量刑前调查报告中建议从重加重处罚少数族裔的倾向。在美国，量刑法官很大程度上会依赖保释官的意见作出量刑决策，一方面造成了辩护律师的意见很难被采纳；另一方面保释官也往往以“量刑专家”自居，滥用其对量刑法官的影响力，将偏见引入案件量刑之中。有外国学者注意到一种现象，如果法官看完量刑前调查报告后，倾向于作出不同的量刑决定，保释官主管或者监管保释官就会给法官提供未披露的额外信息，以至于地区检察官建议的量刑几乎总是达到或者接近法定刑期上限。^{〔41〕}只要这种重刑倾向没有被识别出来，就可能进入训练算法的基础数据库，人工智能工具在不断的机器学习之中重复、固化甚至放大对少数族裔的偏见，影响未来的案件裁判。在卢米斯案中，有学者认为，认知基础数据的偏见（cognitive biases supporting the reliance）是因为裁判者更信赖看似客观的科学证据。罪犯矫正分析系统作为多年、多种专家经验的集合，比个别专家可信，更比被追诉人自己的陈述辩解更有说服力。对于美国少数族裔保释风险的实证分析，也验证了这种偏见的固化效应。算法本身并不能提供评估保释风险的标准，保释风险的标准是通过以前的裁判数据通过算法的机器学习而得来的，但是以前的执法数据中，美国司法警察趋向于对少数族裔尽量不保释，在此基础上形成的保释风险算法自然也趋向于不保释少数族裔的美国人。

（二）算法偏见归责的障碍——喧宾夺主的辅助者

算法偏见追责的虚无状态，需要归咎于算法对“主体性”的侵蚀。通过司法制度发展历程的分析可以发现，算法偏见存在于也发展于司法网络化、智能化的进程之中，体现为数据技术减轻裁判负担，从文件扫描、语音输入到类案数据搜索推荐、文书模板生成，从司法决策外围走向核心环节，减轻司法工作量，同时也以“技术”替换“人工”，侵蚀着司法裁判的“主体性”。为了

〔40〕 张玉洁：《智能量刑算法的司法适用：逻辑、难题与程序法回应》，载《东方法学》2021年第3期，第195页。

〔41〕 文章作者举例称，“一起案件中，保释官在《审前量刑评估报告》正文之外，私下向法官提交以下全部量刑情节信息，都绕过了律师的质证与辩护，包括：（1）被告人曾经很长时间是主要的毒品卖家；（2）被告人引诱其妹妹吸毒；（3）被告人定罪后还在贩毒；（4）被告人涉嫌其他犯罪，如预谋抢劫并杀害一对老夫妇；（5）被告人是共犯的主脑；（6）被告人通过过去违法行为大量牟利；（7）被告人对警察、保释官不合作”。Stephen A. Fennel & William N. Hall, *Due Process at Sentencing: An Empirical and Legal Analysis of the Disclosure Pre-sentence Reports in Federal Courts*, 93 Harvard Law Review 1613, 1670 (1980).

恰当配置“人”与“算法”因素的关系，关键是要解决人的“主体性”保障问题。有学者提出了司法智能辅助系统带来的“法官主体性消隐危机”⁽⁴²⁾，法官的主体性不仅在于裁判文书落款的法官姓名，更在于实质的司法权衡过程，即权衡正反面事实证据的可信度，斟酌选择裁判的价值平衡点，一旦有了算法辅助与参考，则“人”的深入独立思考、无限趋于公正的内心衡量过程可能被消解，类案参考建议的清晰数理逻辑要驳斥并不容易，价值衡量的法官自主思考证成合理性也不容易，长此以往算法可能喧宾夺主，裁判中“人”的因素趋向退让，“机器”的因素因算法运用而拓展，直至全社会面临这样一个问题——我们需要的公正裁判是非人的、机器的因素占主流的吗？在刑事司法运用中，“人”的因素、“机器”的因素之间此消彼长的动态关系值得关注，内含固有偏见、放大验证偏见、循环结构性偏见的智能辅助系统是否及何时会偏离“辅助”的轨道，值得忧虑。

算法偏见还源于监管责任的错位，算法被寄予促成司法体系规范化的厚望，能够减少裁判者个人的偶然性因素对裁判结果的影响，但是这样的期望往往一方面规范裁判者的个案裁量，另一方面也带来了“谁来规范规范者”的困境。在卢米斯案中暴露的倾向是，法庭被要求承担所适用的裁判辅助工具偏见审查的职责，虽然名义上算法仅仅是“辅助”工具，是司法机关查阅参考智能辅助系统的结果，但实际上类案审查已经成为一项必经程序，而且如何“辅助”实际是各个裁判者都须面临的难题。这个对算法的“辅助”定位，实际上是让裁判者承担纠正算法偏见的责任，但是显然法庭无法理解算法原理，更难以遏制偏见入侵，模糊的裁判者与大数据系统之间的责任分配关系，更容易固化偏见，纠偏机制也付之阙如。况且，即使法官希望纠正，也存在着算法数据运用的悖论难题，“法官改判是对智能量刑算法‘量化规范逻辑’和‘经验规范逻辑’的否定，而且改判后的反例也将引发智能量刑算法‘归纳推理方式’的司法适用危机”。⁽⁴³⁾ 原一审判决因为二审或再审改判，可能出现算法生成后难以自恰的逻辑难题。

无人负责的成因是算法技术应用的超前性，算法的刑事运用是在“技术先行”的思维指引下发展起来的，算法的运用实务远领先于算法规范，目前尚无对算法责任与安全风险的有效规范。之所以出现无人负责的情形，是因为算法技术让司法裁判人员的主体性受到侵蚀。一定程度上，司法人员的主体性被“辅助”所喧宾夺主，主要是指人的思维功能，尤其是在司法裁判中的“判断”功能实际上被程序算法所替代。虽然不是全部的判断——事实、价值、法律之中的判断，事实判断一定程度上被接管、法律判断的逻辑大小前提被接管，价值衡量则几乎仍是处于人类思维的保留地，暂时无法染指，但是对于“人”沦为客体，被机器或看不见的程序所审判、所决定命运的恐惧感，深深摄住了人心。

在算法面前，人的思维能力包括记忆力和计算速度都有局限性，难以与算法所依托的计算机

(42) 杜雨钊、唐力：《人工智能辅助司法裁判的正当性边界与制度约束》，载《新疆大学学报(哲学社会科学版)》2025年第3期，第107页。

(43) 张玉洁：《智能量刑算法的司法适用：逻辑、难题与程序法回应》，载《东方法学》2021年第3期，第195页。

系统媲美存储量与计算速度，在数据占有、处理能力上，人类并不占优势。面对强大的算法，人类处于弱势地位，算法已经是一种新兴的权力。刑事司法决策中的算法运用，如果没有限制，可能沦为“恣意的权力”，尤其要约束算法偏见，避免权力有意无意地被滥用。

（三）介入算法偏见纠正的障碍——司法决策过程不可拆解

量刑辅助决策系统依赖计算机技术语言、加工基础数据库，如果算法无法清晰地解释关键步骤，迎合了量刑决策者对评估手段可靠性的依赖，就放任了评估工具依赖前提中陈旧的社会成见。^{〔44〕} 算法是根据“过去经验”概括“未来可能性”的规律。在人工智能发展领域中，不可知论的观点一直挥之不去。根据英国哲学家、不可知论者休谟的认知分析，人即使妄称已经从过去的经验中得知了物体的本性，那也无济于事。^{〔45〕} 面对同一被追诉人的同一犯罪行为在不同社会背景下，刑事法官可得出的公正判决也是不一样的，为此，抽离社会环境，单纯依赖类案总结却忽视个案判断，则有着刻舟求剑的风险。

算法所依赖的类案判断实质是数据统计，所作的概然性判断本身不能等同于“因果关系”的推理。即便休谟持不可知论，但是其关于归纳推理缺陷的分析是有价值的。在休谟之后，不论是在哲学界还是科学界，人们一向认为在对实际事物的认识中，运用归纳推理所得出的结论只具有概然性，而不具有必然性。^{〔46〕} 尤其是在人工智能应用于刑事司法的情况下，机器学习缺乏“人的意志”因素，更无法区分新旧情况的差异，以及其他低概率事件发生的可能性。算法本身是没有判断目的和意义的，只是依照人类编写的代码运行程序。^{〔47〕} 算法即使得出概然性的判断结果也只能还原统计过程，对个案价值衡量却是不可解释的。在计算机专业语言的门槛阻挡之下，法官无法对具体的决策过程进行审查，“决策相对人因对决策相关信息缺乏基本了解而难以拥有应对与反制能力，在信息获取和决策判断能力方面均居劣势”。^{〔48〕} 在卢米斯案件中，作为支持采用罪犯矫正分析的重要依据之一，法官认为量刑前调查报告虽然运用了罪犯矫正分析系统的评分，但任何量刑前调查报告提交给量刑法庭前，都必须附上适用限制的提示，而且提示必须告知量刑法庭，有研究表明，风险评估可能不准确，存在不成比例地高估少数族裔被告人再犯风险的可能性。^{〔49〕} 作为一项普遍告知的量刑程序书面建议，该内容实际上承认了罪犯矫正分析系统对少数族裔可能存在算法偏见。终审判决法庭尽管强调罪犯矫正分析系统结果最终交给法官自由裁量，但是法官并没有能力实质审查分析结果的准确性。卢米斯案判决回避了算法是否能对个案作出准确评估这

〔44〕 Kia Rahnama, *Science and Ethics of Algorithms in the Courtroom*, 1 University of Illinois Journal of Law, Technology & Policy 174, 179 (2019).

〔45〕 [英] 大卫·休谟：《人力理智研究》，周晓亮译，中国法制出版社 2011 年版，第 31 页。

〔46〕 阎吉达：《休谟思想研究》，上海人民出版社 2019 年版，第 227 页。

〔47〕 蒋佳妮、堵文瑜：《促进人工智能发展的法律与伦理规范》，科学技术文献出版社 2020 年版，第 63 页。

〔48〕 蔡星月：《算法决策权的异化及其矫正》，载《政法论坛》2021 年第 5 期，第 27 页。

〔49〕 原文为：Some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism. See *State v. Loomis* 881 N.W. 2d 749 (Wis. 2016), <https://cases.justia.com/wisconsin/supreme-court/2016-2015ap000157-cr.pdf?ts=1468415026>.

一关键问题。如果算法本质只是统计数据,无法评估个案,同时又因为法官不理解上述系统的原理,即使知道其分析结果可能不准确,也不必然能够识别出具体的不准确因素,实际上案件就没有合格的法官予以审判,获得合格法官审判的权利因为算法而被剥夺。换言之,“不透明性源于统计模型或源代码的专有性特征,而这种特征是受到商业秘密相关法律保护的”。⁽⁵⁰⁾ 如果算法偏见存在,则算法的不可解释性会造成一种失控状态,即法官无法监督消除偏见。

在“辅助”的定位之下,智能辅助系统介入了司法决策形成过程,通过制作“类案检索报告”的方式,从类比思维、相似点对比的角度实际开展价值衡量。上述分析本应当以人的思考为主要内容,算法则以事前预设、权重赋值的方式简化处理了价值权衡难题,可能存在僵化适用、机械主义的趋向,无法适应极具复杂性的刑事司法裁判工作。“价值权衡”始终被认为是人类思维的最后自留地。算法参与司法决策,也需要受到程序的约束。对算法而言,算法设计、数据练习、复杂神经系统的机器学习都需要增强算法透明性、为制衡算法偏见而增强被追诉人知情与抗辩的权利能力,有学者总结“‘人’的逻辑不能被‘机器’的逻辑所取代。这是司法的尊严的要求,更是人的尊严的要求”。⁽⁵¹⁾ 只有在法官不被复杂机器语言所束缚,法官绝对主体性得以确立的情况下,算法的辅助地位才会真正确立。

算法辅助司法决策,与传统的司法推理有着典型的不同,是二进制的代码推理、数理逻辑,与司法的演绎或类比推理、法律逻辑完全不同,尽管算法多用于寻找类案,提供参考的量刑建议,但是类案查找方式并非完全等同于人类的“类比推理”,尤其是在类比技术上,存在着显著的差异。类案检索也存在严重滞后性,在司法裁判这一应用场景中,对于待决案件被裁决的这一时刻点而言,同类案件的历史数据是有限的。⁽⁵²⁾ 为此,类案检索至少涉及信息获取、思维工具转换与误差核验纠正三个环节,有学者总结风险包括法律层面的数据错误、数据样本不全面、算法模型不准确不公正,以及程序不透明。⁽⁵³⁾ 其具体对应了以下两方面的问题:一方面是数据提取,即算法生成所依赖的基础数据,数据清洗中需要加入“算法偏见”排除的要求,而且算法偏见需要数据库的完整准确,为此需要遵循严格的社会抽样方法开展数据库提取工作,并且每一步都具有可回溯性;另一方面是算法运用,算法运用要有明确的目的,比如“类案检索”应当匹配到某一个具体量刑情节分支上,而不能笼统、宽泛地“类比”,因为这样的算法推送实际上是对“类比”所采取适用标准的掌控,构成了价值衡量环节的放弃,一定程度上是怠于履行审查的“自由心证”责任。围绕“自由心证”的核心环节,定罪量刑的主要方法仍是依赖“人”的价值权衡及自由心证,而不是“机器”的数理统计与关联分析,为此需要确定智能系统的“辅助”边界,即但凡是

(50) 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》,载《东方法学》2020年第3期,第81页。

(51) 雷磊:《司法人工智能能否实现司法公正?》,载《政法论丛》2022年第4期,第81页。

(52) 杨宇冠、张沈镔:《指导性案例的功能与实施途径研究》,载熊秋红主编:《诉讼法学研究》第26卷,中国检察出版社2022年版,第21页。

(53) 王燃:《大数据证明的机理及可靠性探究》,载《法学家》2022年第3期,第62-65页。

量刑阶段情节权重的考量,主要仍依赖法官,并且法官对此负有职业责任,同时作为参考的类案,并非笼统地都需要参考,而是有必要运用类比推理找到真正具有可参照性的“类案”。

(四) 不脱离演绎推理的算法思路——类案分析的思维僵化

有学者概括,人工智能在司法领域中的使用,本质上还是停留在“自动化司法”的层面,基于知识图谱式算法而缺少定量实证的方法,规范取向而非经验模式⁽⁵⁴⁾,体现为思维方式仍以单纯的法条分析三段论逻辑为主线,无法融入实务中法官裁判的经验,虽然算法的应用是以“类案适用”的面目出现,但可惜的是,现有人工智能的逻辑仍是以符合法律规则为主要方向的,推理方式仍然是演绎推理,并非类比推理,难以体现裁判规则的实质适用效果。对于类案偏见的预防,典型方式并不是“三段论”式的推理,因为核心是放弃了对于类案查找最为核心的“相似性”查找,类比推理过程包括寻找待决案件与类案的相似点予以对比,并且衡量所对比相似点在整个案件中的重要性,“司法办案人员还需根据可能适用的裁判规则,选择对比点,确认哪些对比点重要,哪些不影响法律适用”。⁽⁵⁵⁾ 类比推理的核心价值是放弃了线性推理,如果算法辅助系统仅仅是在“类三段论”方式下进行推理,实际上束缚了类案分析的思维,陷入了南辕北辙的境地。可见,核心要求应当是锁定类似特征,即便作了类比,也应当反向查找总结以往法官的裁判,构成类案有几项关键因素,关键因素又各占多少比重,以这样一个“反向工程”的类似度查找工作,为我国破解类比思维的内核提供分析工具。

现有类案推送算法思维的参照系过于简单,不应当只是或者不主要是“三段论”式的思维,尽管该方式可以最大程度地追求精确性,但是演绎推理的视野狭窄、受限性在以往的司法实务中早有体现。算法追寻的匹配裁判思维,不应只是匹配现状,更应匹配制度发展的需求,即指导性案例“融入裁判思维”的算法推送模式。运用智能辅助系统的核心目的,不是法条解释,或者逻辑涵摄,该部分并非数量逻辑、二进制推理所擅长的内容,相反,多维度案例相似度分析,反而可能成为一种已有技术上更贴切的算法发展思路。作为总体设计方案的引子,有学者建议为了预防算法歧视破坏刑事审判,一是重视形式审查确保初始算法设计公正;二是注重刑事审判算法自我纠错能力;三是增强刑事审判系统算法的透明。⁽⁵⁶⁾ 在此,强调的是辅助算法与法官主导诉讼程序之间的主从关系,通过诉讼程序自有的正当法律程序矫正算法偏见,而不是放弃程序正当性原则,过分让位于算法的神秘性与封闭、机械式运用。

算法偏见不源于算法黑箱,算法透明性策略也不能完全适应算法偏见的治理需要,即便算法透明也解决不了权力结构问题。“普通公众无技术能力来行使对抗几乎无处不在的算法治理的权

(54) 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7页。

(55) 杨宇冠、张沈镏:《指导性案例的功能与实施途径研究》,载熊秋红主编:《诉讼法学研究》第26卷,中国检察出版社2022年版,第21-22页。

(56) 杨玉晓:《人工智能算法歧视刑法规制路径研究》,载《法律适用》2023年第4期,第94页。

利，难以防范算法权力异化。”⁽⁵⁷⁾ 技术进步的发展方向可能不受控制。根据“柯林格理奇困境理论”⁽⁵⁸⁾，“在技术萌芽期，社会无法准确预测一项新技术的不利后果，进而任由新技术发展，直至负面后果自然显现。但是此时，新技术已经深深地扎根于社会、经济结构之中，社会控制已经很困难了”⁽⁵⁹⁾。智能辅助决策系统有利于提升司法效率、增强统一法律标准的类案适用技术，但是不能对此完全放任并且假定其具有稳定的追求公正价值导向。例如有学者建议，“具体到个人信息保护领域，可考虑将资金用于完善公共事宜的算法和数据监管技术等方面”。⁽⁶⁰⁾ 上述算法监管的重要问题，就是尽早植入克服算法偏见的标准，避免该领域算法技术过分发达后的“柯林格理奇困境”。当算法辅助决策技术用于司法裁判的初期，社会可能难以预见算法偏见的危害，认为没必要调整，这正如罪犯矫正分析系统运用得很好却歧视少数族裔的问题，当发现其危害的时候，该系统已经深深扎根于美国量刑程序之中，再要规制，就需要更改该系统的算法与基础数据，这可能需要耗费大量资金，因而这一困境真实而难以处理。为了合理利用大数据工具，欧洲国家确立了人工智能在司法系统中运用的五项原则——尊重基本权利、不歧视、保障质量与安全、透明无偏倚和公正，以及“在用户控制下”适用的原则。⁽⁶¹⁾ 其中，三项原则与算法偏见直接相关，“不歧视、透明无偏倚和公正，在用户控制下适用”都一定程度上涉及发现与纠正算法偏见的问题。

三、算法偏见的刑事正当程序矫正

值得关注的是，自 2021 年起，我国最高人民检察院在全国范围内推行“逮捕社会危险性评估规范化”专项试点，各地纷纷尝试将影响逮捕社会危险性判断因素予以细化量化，实际上正在为后续的在规范化社会危险性评估基础上引入算法模型，为批准决定逮捕、羁押必要性审查等环节人工智能运用做好了铺垫。在这一情境之下，算法偏见的应对研究，有着刑事司法保障的现实意义。有学者提出，为了有效约束人工智能运用，防范算法歧视泛滥，需要发挥司法人员本身的作用，“缺乏‘自由心证’矫正的人工智能极易陷入‘法定证据制度’的误区”。⁽⁶²⁾ 我国司法智能辅助系统的发展关键，在于保障庞大的数字机械逻辑面前的个人权利，甚至“允许信息主体在案件审结后的特定情形下以申请的方式要求封存或删除涉诉个人信息，从而减弱因人工智能技术运用

〔57〕 梁志文：《论算法排他权：破除算法偏见的路径选择》，载《政治与法律》2020 年第 8 期，第 96 页。

〔58〕 英国阿斯顿大学教授 David Collingridge 于 1980 年提出“控制困境”理论，用于描述技术发展与社会控制的关系，被称为“Collingridge Dilemma”。

〔59〕 Adam Thierer, *The Pacing Problem, the Collingridge Dilemma & Technological Determinism*, <https://techliberation.com/2018/08/16/the-pacing-problem-the-collingridge-dilemma-technological-determinism/>.

〔60〕 陶加培：《个人信息保护检察公益诉讼的路径优化：以多元协同理念为核心》，载《华东政法大学学报》2024 年第 5 期，第 91 页。

〔61〕 European Commission for the Efficiency of Justice: *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment*, Council of Europe, p. 5–12.

〔62〕 程凡卿：《我国司法人工智能建设的问题与应对》，载《东方法学》2018 年第 3 期，第 125 页。

带来的国家强力控制个人信息的风险”。⁽⁶³⁾ 根据我国司法体制对算法运用于刑事司法决策的定位,辅助系统的功能只能是“辅助性的”。司法实务中的算法适用有超出“辅助”定位的趋势,导致了司法人员的主体性日益受侵蚀,主从关系也存在一定程度的错乱。为规范司法大数据运用,一方面技术赋权,对司法责任主体明确权力、限定职责,对被追诉人赋予权利,提供救济与知情通知等权利行使的保障;另一方面予以程序规制,在算法偏见的发现、纠正、救济各个环节设置保障措施,避免陷入算法偏见的陷阱之中。

(一)“技术祛魅”下的算法偏见发现

算法偏见难以被发现在于其技术性的外衣,以复杂的数量逻辑为隐蔽性的保障。算法也有了数字时代“拜物教”的雏形,呈现出一种有意神秘化的色彩。发现算法偏见的第一步是技术祛魅,根据算法生成的基本步骤,整理可能暗藏认识偏差、传统与新生偏见交织的司法认知与决策环节。算法作为一种技术产品而生成,又继续作为技术工具而产生辅助量刑成果,基于这一属性,笔者建议通过监控设计指标、审查基础数据、允许程序异议三方面的手段,发现与监控算法偏见。

1. 落实算法设计者责任而发现偏见

在数字时代,算法运用本身就是一种创设权力的方式。司法决策中算法权力滥用应当被追责。值得关注的是,算法设计者处于前端,开展算法开发、训练与运营,由于其实质介入司法决策,却游离于现有的司法责任约束之外,亟须予以专门规范,避免其滥用算法设计权力而引入算法偏见。

在算法设计阶段,有学者建议嵌入反歧视要求,人工智能的开发首先须具备正当目的,不得被设计用于执行歧视性任务;训练算法模型的数据集须具备均衡性;在技术效果近似时,优先选择更利于实现平等的方案。⁽⁶⁴⁾ 刑事司法的算法开发,也需要遵守“算法解释”的要求,即在算法适用于具体刑事个案之前,司法适用的算法基于其重大公共利益影响,对其算法设计问卷,包括算法设计源代码、基础训练数据集、算法训练方法应当有一定程度的收集与整理,以供事后审查与公众监督;算法设计者对算法可能存在的固有偏见、引发偏见的适用场景,予以充分的事前测验与披露记录,提交给专门的司法决策辅助算法审查机构予以核查,避免因为“算法黑箱”的交付,对刑事司法决策过程埋下隐患而不自知。算法设计的技术指标合理性应当经过充分论证,“因为量刑情节并非仅仅是一种客观行为,还涉及到司法人员的主观认定问题,不可避免地会带有司法人员的主观判断”。⁽⁶⁵⁾ 算法偏见存在于参数选取的过程,亟须在法学领域引入更精深的量化研究,细化衡量类案类比中各类因素哪些被选用、哪些被排除以及各自权重如何设置等问题,以量化研究指导减少算法设计过程的指标选取的恣意与指标赋值的主观性。

(63) 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,第693页。

(64) 李成:《人工智能歧视的法律治理》,载《中国法学》2021年第2期,第145页。

(65) 高通:《逮捕社会危险性量化评估研究——以自动化决策与算法规制为视角》,载《北方法学》2021年第6期,第137页。

2. 构建围绕算法开发、运行的算法偏见伦理审查机制

在司法决策的场景中,越来越多的算法有赖于专门审查、发现与纠正算法偏见,为此有必要由具备专门能力、配备专门人员的机构开展上述工作,笔者建议由常设主体承担这一工作,即设立算法伦理审查机构。“在任何时候,机器都不可能完全代替人类作出主观决策和价值判断。”⁽⁶⁶⁾ 算法偏见也需要通过人工审查,为此可以采用算法伦理审查方式发现并纠正算法偏见。

随着算法在检察机关、司法机关日益普遍运用,有必要专门设置算法伦理审查委员会审查与处理算法偏见问题,该委员会应当包含算法技术与司法技术两方面的专家,并且对算法设计、试运行与日常投诉问题予以处理。算法伦理审查委员会应当设置多步骤的审查程序:一是初步审查。初步审查由算法设计方提出,初步审查应当在算法正式投入前进行,为了保障算法的审查深度,算法设计方应当披露算法伦理审查委员会行使职责所需要的全部数据,对其提问提供必要的解释。二是试用审查。初步审查由司法机关人员或利益相关方向算法伦理审查委员会提出审查请求。司法机关明确司法人员的算法偏见防范责任,对算法设计者提出完善要求。三是算法偏见的日常投诉处理。算法伦理审查委员会不干扰司法裁判,不介入具体个案的事实裁判,但是考虑算法偏见的发现与纠正难度,算法伦理审查委员会还可以接受个人对个案中可能存在算法偏见问题的日常投诉,作为一种常设机制,算法适用各环节涉及的主体,包括被追诉人,都可以投诉可能影响个案公正但在司法审查中法官无法发现、无法了解其解释原理的算法问题,要求审查纠正。

3. 审查算法基础数据而发现偏见

筛选、排除基础数据库的固有偏见,“犯罪预测类算法在投入使用前,应充分在不同地区和领域测试,收取达到规定数量的数据集,并测试其在打击犯罪中的实际作用”。⁽⁶⁷⁾ 近年来,审查筛选类案已经非常深入,有了很好的数据筛查经验基础,使用 AI 一定要告知其风险,“充分提示 AI 算法的固有缺陷和使用风险,披露 AI 模型的主要参数以及法律推理的主要逻辑,以满足算法审计需求及透明度要求”。⁽⁶⁸⁾ 然而,“算法偏见”是非常突出的一个问题,与“算法黑箱”并非同一个问题。

现有研究中很多将黑箱问题与算法歧视混同,这实际上对于二者的理解存在重大偏差。因为算法歧视不必然是来自算法的有意识的封闭性,而是在于无意识的、思维结构的缺陷,以及算法逻辑对价值衡量的忽略,人与算法的思维不匹配的问题,不可避免地忽略了僵化而未能解决的算法适用问题。“透明性、可解释性等原则”并非算法偏见的出路,偏见需要“匹配”司法立场、适应刑事审判期待其实现的功能,同时保障被追诉人乃至被害人的辩护权与申诉机会,算法歧视或偏见的核心是算法助长“偏听偏信”的效果。为规范算法偏见,司法机关应当确立算法适用的范围与案件类型。根据最高人民法院对“类案检索”的要求,“除指导性案例以外,优先检索近三年

〔66〕 吴雨豪:《量刑自由裁量权的边界:集体经验、个体决策与偏差识别》,载《法学研究》2021年第6期,第128页。

〔67〕 张凌寒:《算法规制的迭代与革新》,载《法学论坛》2019年第2期,第24页。

〔68〕 李晓楠:《可信赖 AI 司法:意义、挑战及治理应对》,载《法学论坛》2020年第4期,第126页。

的案例或者案件”。⁽⁶⁹⁾有学者讨论数字在刑事诉讼中运用的平等性原则，注重“这些要素主要包括四个方面：接入性、可负担性、能力以及内容相关性”。⁽⁷⁰⁾如果算法所涉的案例不能反映近些年的裁判趋势，在急剧变化的司法领域，不应当简单地套用类案制度下的算法约束规则，避免出现错误。

4. 允许算法程序异议而发现偏见

在拟采用算法的刑事案件中，被追诉人应当有权提前知悉并自行决定是否拒绝适用算法工具，从而保障算法工具的适用以“知情同意”为前提，避免由机器在违背被追诉人意愿的前提下实质作出影响个人切身利益等的司法裁判，因为程序公正的重要前提是参与性，即作出对被追诉人不利裁判的过程允许被追诉人按照个人意愿参与其中；被追诉人享有算法缺陷的质证权，从而保障可能遭遇算法偏见的主体，能够主张自己的程序权利，对算法问题提出异议。

有学者从社会学角度分析认为，如果局限于人与机器互动的环境之中，人总是处于等待被机器系统识别的被动状态，在面对复杂的算法时，司法人员并非操控主体而只是被提供算法评估结果的对象，“在一定意义上，是人类被自动化了”。⁽⁷¹⁾在对资本主义下社会控制方式的分析中，韦伯要求排除个人好恶而个案衡量地处理事务⁽⁷²⁾，算法固然可以排除人类法官的主观倾向，但不能丢失人类司法最低限度的对个案公正的追求，即算法运用可以排除对被追诉人个人的好恶因素，却不能丢失对个案犯罪行为本身的具体衡量。为了发现算法偏见，司法体系仍须对算法偏见保持足够警惕，正视人工智能的发展现状，更多地发挥司法人员的主体性与审查能力，发现算法偏见。

出于刑事政策考虑，司法系统整体推行智能辅助系统。算法偏见之所以难以被发现，是因为算法的技术封闭性。算法偏差的发现，需要从算法本身的设计开始，因为算法也是一种技术产品，其生成过程本身不可避免地包含偏见植入、叠加的过程。在算法应用中，一大部分的算法偏见预防工作需要借助对算法设计、生成阶段的约束来实现。算法偏见防范的核心是从技术角度提供约束，涉及了技术、各方行为与救济保障。以算法的透明公开为审查发现算法偏见的前序步骤，规制对象需要可回溯式的“算法设计者必须遵守的伦理规范”⁽⁷³⁾，司法决策的算法辅助中，算法的设计者和司法人员都应当受到规制，开展算法说明、算法筛查、算法第三方监管和算法风险评估⁽⁷⁴⁾，约束算法运用。

(69) 《最高人民法院关于统一法律适用加强类案检索的指导意见（试行）》（法发〔2020〕24号），第4条第2款。

(70) 裴炜：《共建共治共享理念下数字社会治理的多主体协同》，载《数字法治》2023年第2期，第22页。

(71) 原文为：Instead of making machines augment human intelligence, the coding elite reorganizes social activities to make humans support the operation of computers. It is thus humans who are, in a certain sense, automated. See Jenna Burrell, *Marion Fourcade: The Society of Algorithms*, 47 Annual Review of Sociology 213, 227 (2021).

(72) [德] 马克斯·韦伯：《支配社会学》，康乐、简惠美译，广西师范大学出版社2010年版，第46-47页。

(73) 郑戈：《人工智能与法律的未来》，载《探索与争鸣》2017年第10期，第83页。

(74) 杨玉晓：《人工智能算法歧视刑法规制路径研究》，载《法律适用》2023年第4期，第92-93页。

（二）匹配量刑决策的算法偏见纠正

回归最基础的数据处理，有学者提出注意保障数据质量的基本原则。“如果不能保证刑事诉讼中数据处理时的质量，亦无法保证其准确和完整，即便成为了诉讼证据，在审判中也会被排除，不能成为认定案件事实的依据。”^{〔75〕} 实务中，智能辅助系统面临“简易案件不需要，复杂案件不敢用”^{〔76〕} 的状况，这是因为司法审查未能以“整体主义”看待全案的量刑证据，对量刑证据审查停留于对定罪证据笼统审查、在量刑裁量中照单全收的阶段，并未细分有别于定罪标准的量刑证据“优势证据证明标准”，也未适用“存疑时有利于被追诉人”的量刑证据审查原则，导致了量刑程序没有从定罪程序中分离出来，量刑证据的认定流于“定性审查”后简单的整体确认。

1. 提供“量刑司法决策”思维结构的基本功能

考虑算法偏见的存在，人工智能的辅助量刑系统应当定位于辅助量刑可视化与可解释性，提供“量刑司法决策”思维结构的基本功能。为了精准确定量刑情节，有必要适用更为精细、准确的量刑事实认定方式，这是旧有定罪证据审查思路所不能涵盖的，也是现有认罪认罚从宽量刑建议下，法检部门试图通过清晰化量刑过程、对量刑建议的形成达成一致而共同追求的，尤其借用智能辅助系统搭建相对稳定的量刑计算方式，实际上承担起了探索清晰化量刑思路的作用。

尽管一直有担忧算法可能侵蚀人类主体性，但算法日益普遍的运用是不争的事实，算法运用价值也毋庸置疑，因此与其盲从算法的科学证据表征，不如运用算法证据伴随司法推理与证明的过程，引向实质的证明辅助功能。^{〔77〕} 我国的 206 系统、检察机关办案法律监督模型的大量发展，也与量刑精细化实践需求相呼应，跨越算法偏见的唯一方法，恰恰是更加充分地拓宽算法适用基础，更规范地适用算法。智能辅助系统的发展，也只有与我国司法证明进一步区分定罪证明与量刑证明，形成更加规范、可解释的量刑情节评判计算公式，并且从整体角度衡量量刑是否公正，才能更加匹配司法审查需要。作为程序保障，司法责任制的“主体性地位”根据责任分配原则而保持不动摇，让审理者裁判，智能系统辅助提供量刑司法决策的“思维框架”，便于裁判者以“可视化”的方式审查量刑准确程度，并且以“可解释”的计算过程表现出来。

2. 寻找确定裁量规范化基准的准确理论

有学者分析，在未完全引入人工智能工具的情况下，逮捕与羁押必要性审查工作存在司法体制考核无动力、审查程序繁琐复杂的问题，“‘逮捕中心主义’的观念没有彻底转变”。^{〔78〕} 这构成

〔75〕 郭烁、杨默涵：《受限、契合与独立：论刑事诉讼数据处理原则》，载《北京航空航天大学学报（社会科学版）》2022 年第 4 期，第 35 页。

〔76〕 谢澍：《人工智能如何“无偏见”地助力刑事司法——由“证据指引”转向“证明辅助”》，载《法律科学（西北政法学报）》2020 年第 5 期，第 111 页。

〔77〕 Edwina L. Rissland, *Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning*, 99 The Yale Law Journal 1957, 1980 (1957).

〔78〕 谢小剑、肖宪涛：《我国审前未决羁押制度的实证研究——以中部某省 5 县区为样本》，载《北方法学》2024 年第 1 期，第 122 页。

了社会危险性评估系统开发中基础数据的固有偏见。进一步而言,有学者认为其定量分析程度显然不足,“我国更青睐于通过定性研究方法来确定,这也使得主观推断、猜测等对我国量化评估模型的影响更大”。⁽⁷⁹⁾从长远发展来看,算法偏见的原因在于规范化的基准问题。以量刑规范化为例,虽然在推行量刑智能辅助系统之前,学界就对量刑规范化基准问题有过讨论:中线论、分格论、形势论、主要因素说、重心论。⁽⁸⁰⁾但是现有量刑用的智能辅助系统,既没有专门公开其确定基准的原则,也没有论证其基准的合理性。有学者在研究中提倡通过集体经验识别个体决策偏差,运用多线层模型(multilevel models)确认量刑偏差水平,以量刑结果偏离集体经验达到1个标准差为阈值,在每个案件类型中,大约有22%~28%的法官进入量刑偏差识别库,以2个标准差为阈值,比例为6%~9%。⁽⁸¹⁾该方法以“正态分布与标准差”作为基本要素,区分了偏差区域,基本遵循了“分格论”的思路设定基准,一定程度上弥合了上述基准方法的争议,但是基准的争议远没有得到解决。相反,因为算法可能作为影响大批案件处理的智能辅助工具,基准问题的研究价值更加凸显。算法运用应当具备牢固根基,基础研究不能止步于此。由于规范化基准是交叉学科问题,统计学与刑事法学的学者应当加强合作,为司法中的算法运用提供牢固的理论基础支撑。

3. 限制算法结果运用的精确程度

为了减少算法偏见的影响,司法人员应当更大程度地限制对算法结论的依赖,甚至有意限缩大数据分析结论的效力,将“定量”精确值仍转化为“定性”区间适用,包容算法偏见的误差。罪犯矫正分析系统往往利用分类加赋值的方式,给出具体危险性打分,虽然罪犯矫正分析系统能够给出再犯危险性的具体得分(1—10分),有“定量”的精确性表象,但美国在运用中仍有所保留,在罪犯矫正分析系统得分基础上再分“(1—4)低,(5—8)中,(8—10)高”的再犯区间,真正适用的结论仍然仅是“定性”区间,也是一定程度地保留司法裁量空间。在允许该系统校准偏差的同时,法官仅参考适用其意见,以司法人员的价值衡量限缩算法偏见的影响。有学者将算法的生成过程分为八个步骤,包括定义问题、数据收集、数据清洗、数据简要审查、数据分割、模型选择、模型训练以及结果评估⁽⁸²⁾,算法偏见来源于设计者的主观态度代入、用于生成算法的数据本身的不周延性或不具有抽样代表的特点,以及在算法训练过程中可能存在的形成过程的各种偏差,为此在降低算法结果适用精度的同时,更需要改善算法本身生成步骤各环节的精细约束与甄别,识别发现并纠正可能的算法偏见。

4. 保留司法人员的自由心证空间

有学者认为,保障算法准确性的核心方式是保障算法匹配于适用场景,“关注算法偏见背后的

(79) 高通:《逮捕社会危险性量化评估研究——以自动化决策与算法规制为视角》,载《北方法学》2021年第6期,第137页。

(80) 周光权:《量刑规范化:可行性与难题》,载《法律适用》2004年第4期,第61页。

(81) 该文的结论是基于47655份2014—2020年北京地区公开的盗窃、诈骗、故意伤害、抢劫、交通肇事罪五类案件基层法院判决分析得出的数据分析结果。参见吴雨豪:《量刑自由裁量权的边界:集体经验、个体决策与偏差识别》,载《法学研究》2021年第6期,第112、118、123页。

(82) 刘友华:《算法偏见及其规制路径研究》,载《法学杂志》2019年第6期,第60—61页。

深层原因，注意识别替代性变量、选择多元化训练数据集，审查模型偏见性以避免陷入恶性循环机制”。⁽⁸³⁾在量刑辅助场景中，司法人员确认量刑事实的方式是自由心证，那么算法运用也需要适应这一场景特点。“不能一味对由算法识别出的量刑偏离作否定性评价。相反，算法的识别只是第一步，还需要通过访谈和实地调研等方式，并加入监督者的主观评价，在对量刑偏差进行全面归因解释后，再对法官的刑罚裁量作出公允评价。”⁽⁸⁴⁾司法人员在发现“自动决策”的算法可能出现偏差时，有权利选择不适用已经偏差的算法结论，从而保留裁判者“人”的因素对算法的必要纠正作用。

随着我国进入轻罪时代，认罪认罚从宽制度适用于80%以上的案件，程序提速、技术门槛提高，司法对“人”进行描述，则主要是对“算法”输入数据，以大数据的方式输出，“技术性”提高，减少了人与人的对话，而诉讼中的对话则是辩护的基本方式，智能辅助系统运用在算法决策参考范围内，一定程度上隔绝了辩护的机会。科技进步的同时，也可能带来程序的隔膜与封闭性，尤其是以算法技术最为突出，“算法”的司法程序决策严重影响诉讼程序的沟通过程，对参与诉讼的被告人、被害人均存在偏见的可能性。唯有解绑智能化系统“类案结论”对裁判者自由心证的束缚，科学界定其与“类案推荐”不同时裁判者的说明责任承担方式，才能让裁判者保持主体性，也对刑事量刑决策的主体性承担后果责任。

（三）引入类比推理的算法思路回归

算法偏见固然有多方面的来源，但重要来源是算法发展与应用思路上的偏差——过分强调既往裁判经验的不可突破性，这一点与“类案”参照适用核心的功能不符。类案检索重视法官自由裁量权的运用，在英美法系“遵循先例”的判例传统国家中，法官须慎重衡量已有裁判中的少数意见，注意案例比较背后的价值衡量，分析世易时移之下“原则—例外”的交替演进关系等，但算法运作与之格格不入。算法分析多数以“三段论”推理为主要的司法推理模仿对象，忽略了引入算法分析的核心程序功能，不在于补强法律解释为核心的演绎推理做法，而是更新以“类比推理”为核心的类案检索与统一法律适用分析方法。为了继承与弘扬我国司法裁量的固有优势，同时又推陈出新地发展出结合英美法系与大陆法系优势的指导性案例参考制度，算法需要引入类比思维的审查与分析方法，以供司法实务人员参考。

根据罗尔斯的《正义论》观点，在无知之幕后，任何人都不知道自己未来在社会生活中承担什么样的角色，则这些人共同制定的程序制度才能最大限度地反映程序特点，构建符合程序公正要求的诉讼制度。考虑刑事诉讼既需要维持实体定罪量刑结果的公正，又需要保障程序公正与正当，刑事司法决策过程本身就是一个以“不完全程序正义”为基本预设的程序过程。据此，算法辅助也应当以“不完全程序正义”为预设，以假设所有算法设计与使用的人员都不知道所承担的角色为前提条件，新设有利于被追诉人的算法适用适当性的质疑权，提示告知算法适用的通知责

(83) 王燃：《大数据证明的机理及可靠性探究》，载《法学家》2022年第3期，第71页。

(84) 吴雨豪：《量刑自由裁量权的边界：集体经验、个体决策与偏差识别》，载《法学研究》2021年第6期，第129页。

任,并且在算法设计者无法就个案刑事司法决策而被追责的情况下,要求明确算法偏见可能存在的情形、场景,以及适用禁忌的告知要求;在司法裁判中,则需要畅通算法的质疑、类案运用中的不当性说理;在司法裁判之后,提供完整卷宗便于二审审级监督;如果实际出现了算法妨碍而不是辅助司法裁判的情况,法官有权说明其算法偏差的情形,从而免除不适用“类案报告”的说明义务。⁽⁸⁵⁾

(四) 刑事司法中数据程序权利的算法偏见救济

算法偏见救济,是偏见规制中的关键一环,只有完善了算法偏见受损方的救济,算法偏见责任才能得以落实。“重视可能建构新的裁判规则的罕见的独特性案例。”⁽⁸⁶⁾司法案例较大程度上是非标准化的工作,算法运用的偏见也在于以简单的数量提炼、标准化的方式处理不尽相同的案件。为此,避免算法偏见本质上需要考虑算法处理案件的方式与司法体制匹配问题。算法的计算逻辑,与司法人员的办案逻辑不一致,名义上算法是办案参考,实际上并未服务于司法审判思维,由于嵌入司法审判的程度不尽彻底,法官的主体性难以彰显。算法成为实际决策者,剥夺被追诉人的权利却无人对此承担司法责任。为平稳推行算法适用应做到以下几点。

首先,赋权被追诉人排除特定情形算法适用的申请权。有研究建议将“需要进行价值判断的疑难复杂案件”⁽⁸⁷⁾列入禁止清单。为了避免算法偏见带来的不利影响,司法机关有必要采取算法中立的方式,保障量刑算法的准确性。为了保障算法中立性,智能辅助系统的设计过程应当有更多元主体的参与,尤其是辩护律师的参与。有学者发现,“从当前司法实践来看,非但当事人及其辩护律师极少能参与到这一过程中来,其意见也很难反映到量化评估模型中去”。⁽⁸⁸⁾算法中立首先要保障被追诉人的基本权利,有学者提出,“算法自动化决策对公民权利影响甚巨,却完全规避正当程序”。⁽⁸⁹⁾受到算法影响的被追诉人,只有选择是否允许适用算法工具的权利,具体做法上可以考虑对算法运用本身设定考察质证的指引。例如,有学者建议对经济犯罪可能运用到的抽样取证,在技术分析的思维下考虑其固有科学性的评价,“从统计学角度,由于误差的客观存在,即便是随机抽样调查的结果也不可能是精确值,只能是估计值”。⁽⁹⁰⁾刑事司法领域的算法设计也更需要考虑算法可信度,算法偏见可能存在应予以提前说明,便于实质的有效质证对抗,以发现与排除算法偏见,最大程度地贴近客观真实与实现刑事司法公正。

其次,赋权被追诉人主张追究算法设计者、基础数据提供者责任的权利。从公共决策研究角

〔85〕《最高人民法院关于统一法律适用加强类案检索的指导意见(试行)》第2条规定了类案检索的适用场景,“拟提交专业(主审)法官会议或者审判委员会讨论的”与“缺乏明确裁判规则或者尚未形成统一裁判规则的”;第10条规定了法官对不采用类案的文书说理义务,“公诉机关、案件当事人及其辩护人、诉讼代理人等提交指导性案例作为控(诉)辩理由的,人民法院应当在裁判文书说理中回应是否参照并说明理由;提交其他类案作为控(诉)辩理由的,人民法院可以通过释明等方式予以回应”。

〔86〕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,第115页。

〔87〕蔡立东、郝乐:《司法大数据辅助审判应用限度研究》,载《浙江社会科学》2022年第6期,第61页。

〔88〕高通:《逮捕社会危险性量化评估研究——以自动化决策与算法规制为视角》,载《北方法学》2021年第6期,第139页。

〔89〕张凌寒:《算法规制的迭代与革新》,载《法学论坛》2019年第2期,第25页。

〔90〕陈学权:《我国刑事诉讼中抽样调查研究》,载《中国刑事法杂志》2024年第3期,第59页。

度分析，大数据技术被分为“信息逻辑”和“决策逻辑”两个部分的内容，二者既有互相促进也有互相制约的因素。其中，互相促进在于“信息逻辑”背后是大数据的信息提供者，通过提供更加全面、更加高质量的信息，帮助决策者；“决策逻辑”背后是决策者，通过更好地选择、吸收有价值的信息，作出正确决策。算法设计者隐身于算法背后，不受司法决策责任的约束，但很大程度上实际影响决策过程与结果，一定程度上是算法偏见的产生源头。为此，不懂法律适用的算法设计者应当有所不为。“某种意义上，司法官对不同案例规则有选择适用的权利。”⁽⁹¹⁾在涉及案例复杂价值衡量时，算法设计者不宜站在俯瞰庭审法庭的角度，轻易以所谓类案经验干扰法官的个案衡量，相反要为个案裁量保留空间。因为“信息逻辑”与“决策逻辑”都可能加入自身的目的视角，提供信息或使用信息，验证各自已经持有的观点或目的，二者既有排斥算法偏见的一面，也有可能因为各自先入之见而助长偏见的一面，“大数据赋予数据分析者与决策者能力，都可以在最终决定作出前，作出选择进而影响决定”。⁽⁹²⁾对于量刑司法智能辅助系统而言，算法的设计者、基础数据提供者责任亟须区分，从设计算法、提供信息的角度，算法偏见可能因为设计者的偏差，也可能因为有意识选择后提供基础数据，带来特定的算法偏见。裁判者更多遵从“决策逻辑”，居于决策者的位置，目前按照司法责任制的要求，对案件结果承担责任，但是算法中立不宜仅仅涵盖“决策逻辑”部分，还应包含遵从“信息逻辑”，区分出算法设计者责任、基础数据提供者责任，要求算法运用的各参与主体承担起保障算法价值中立的责任。

最后，提供司法辅助算法的证据开示等程序保障。有学者指出，“算法本身会有一种反馈循环机制（Feedback loops），一旦机器学习的运算结果得到验证反馈，则会强化其模型中的某些变量，进而产生更多的该类结果”。⁽⁹³⁾算法运用尤其需要警惕偏见的反馈循环，针对证据类的算法适用，可以让专家辅助人对算法核心运算机制进行质证，在量刑决策等重大诉讼事项中，保障“人”的因素有效参与到算法运用过程之中，尤其是对被追诉人不利的裁判结果形成过程中，允许被追诉人借助专业力量充分参与其中，保障被追诉人对算法偏见质疑的机会。

On Impact of “Algorithm Bias” on the Due Process of Criminal Justice and Responding Suggestions

ZHANG Shenqie

Abstract: As AI assistant system in criminal justice has been dramatically expanding,

(91) 黄京平：《刑事司法人工智能的负面清单》，载《探索与争鸣》2017年第10期，第88页。

(92) H. G. Van Der Voort, A. J. Klievink, M. Arnaboldi & A. J. Meijer, *Rationality and Politics of Algorithms. Will the Promise of Big Data Survive the Dynamics of Politic Decision Making?*, 36 *Government Information Quarterly* 27, 36 (2019).

(93) 王燃：《大数据证明的机理及可靠性探究》，载《法学家》2022年第3期，第64页。

influence of algorithm bias on criminal justice attracts more attention. Algorithm bias is rooted in the inherent attribute of the non-neutral value of the algorithm, which should not be simply mingled with the algorithm black box problem, and cannot be dealt with in the common regulatory path of “enhancing transparency and attributing responsibility”. Algorithm bias is often found in unconscious thinking frames and blind spots of perspective, and there are special cases such as solidifying, anchoring and confirming bias outside of information asymmetry. In criminal justice, the algorithm bias comes from the rigidity of the thinking of case-by-case analysis. It is suggested to return to due legal procedures, disenchantment of algorithm technology, discovery of algorithm bias, rectification of bias matching the actual thinking process of sentencing decision-making, protection of the ability of the accused to raise objections and procedural rights, introduction of analogical reasoning, so that the application of the algorithm is adapted to the development of the current path of “unified law application” in class case retrieval, and the actual needs of judicial decision-making are more accurately matched.

Keywords: Algorithm Bias; Algorithm Black Box; Class Case Analysis; Due Process of Law; Technological Disenchantment

（责任编辑：陈学权 汪友年）