刑事司法科学与治理
Criminal Justice Science & Governance
2025 年 第 6 卷 第 2 期

人机协同视角下治安处罚裁量辅助系统的技术原理、风险检视与应对策略

常克强

中南财经政法大学刑事司法学院，武汉

摘　要 | 人工智能技术和公安执法的结合势不可挡。本文通过对武汉迪赛威智能科技有限公司的实地调研发现，目前公安执法正从“实体决策”向“实体预测”阶段演进，技术应用也从“专家主义”向“连接主义”转型。然而，学界现有研究多聚焦于法检系统和“专家主义”相关领域，针对公安执法中“连接主义”技术应用的研究尚属空白。以此为切入点，本文首先明晰了技术的原理与构建内涵；其次以相关性与因果性、主观性与客观性、独立性与依赖性为核心矛盾原点，归纳出理论界公认的程序风险、实体风险与算法风险三大风险类型；最后结合法律大模型的新型应用背景，对上述风险再次审视，并以“人机协同”理论为支撑，从应用策略、算法策略与程序策略三个维度提出相应风险应对路径，旨在实现技术优势和执法人员专业优势的有机结合，达成共赢局面。

关键词 | 人机协同理论；计算法学；技术原理；风险检视；应对策略

一、引言

当今社会，人机互动十分频繁。国家高度重视将人工智能技术与政务工作相结合，加强执法裁量智能化建设。在制度层面，我国已实现从智慧政务宏观顶层设计向行政执法人工智能化微观应用的转型，在此基础上，又不断明确与细化相关执法标准。[1] 在科技强国战略背景下，全国各地政府部门与司法机关不断推进人工智能应用落地：以

［1］2017年，国务院印发的《“互联网+政务服务”技术体系建设指南》，其强调要优化审批流程，推广“互联网+政务服务”模式。同年，国务院印发的《新一代人工智能发展规划》强调推进社会治理智能化，加快建设政府服务与决策的人工智能平台。2019年，国务院印发的《国务院办公厅关于全面推行行政执法公示制度执法全过程记录制度重大执法决定法制审核制度的指导意见》强调要加强人工智能技术在行政执法实践中的运用，研究开发行政执法裁量智能辅助信息系统。2021年，国务院印发的《法治政府建设实施纲要（2021—2025年）》强调要积极推进智慧执法，加强信息化技术、装备的配置和应用。2022年，国务院印发的《国务院办公厅关于进一步规范行政裁量权基准制定和管理工作的意见》强调要细化执法标准。

政府部门为例，各地开始使用人工智能辅助审批等程序性任务，也开始辅助进行决策等实体性任务[1]；以司法部门为例，人工智能主要应用于量刑预测、类案推送等任务。[2]但是在法律层面，国内对人工智能的应用仅仅停留在宏观制度层面，未在具体应用领域进行规定。[3]综合来看，其难以满足当前科技辅助执法的实践需要。

从技术角度切入，将技术渗入执法程度进行划分，不难发现公安系统相较于司法系统存在明显落后。当前，国内各地国家机关与人工智能的结合已从“程序优化”逐步迈向“实体决策”，部分司法部门更已进阶至“实体预测”阶段。值得注意的是，公安机关目前还停留在“实体决策”层面[4]，但近年来已开始逐步向“实体预测”方向探索。

从执法人员应用技术的角度切入，学者们提出了质疑。一方面，我国有公开的权威法律数据，如中国裁判文书网，因此具有计算的可行性；另一方面，裁判文书内容比较简单，涵盖范围比较容易受限，大多以裁判结果和证据清单为主。[5]从量刑价值取向来看，决策者可能受到司法潜见等因素影响，但这些无法在裁判文书中体现。从技术转化的角度来看，法律语言本身具有模糊性，无法准确转化为具体代码，可能引发算法歧视[6]和算法黑箱[7]问题。[8]因此，理论界对人工智能辅助决策提出了质疑，归根结底是不相信人工智能能够代替法律决策者的主体地位。

值得注意的是，目前人工智能技术已经从“专家主义”向“连接主义”发展。[9]在“连接主义”的当下，结合法检系统智能量刑的理论研究和实务经验，对新兴的治安处罚裁量辅助系统风险进行检视十分具有必要性。基于理论与实践的哲学思想，实践决定理论，理论反作用于实践，而正确的理论会促进实践的发展。因此，通过再次检视能够

---

［1］甘肃省环境保护厅搭建了环境行政处罚裁量辅助决策系统，其内容主要是建立关于环境违法行为与违法后果的二元数学模型。广州市工商局搭建了综合执法系统，包括了记录，审批的程序裁量和决策的实体裁量。国家税务总局辽宁省税务局搭建了金税三期税收管理系统，其内容主要是自动计算出对于申报、发票、登记等日常征管类处罚结果。

［2］湖北省检察院、安徽省怀宁县检察院、广东省广州市南沙区检察院搭建了智能量刑辅助系统，其功能包括智能量刑预测、法律法规检索、刑事案例检索。浙江省安吉县检察院搭建了“安心量”最佳案件智能辅助系统，主要是针对醉驾案件进行量刑辅助建议。江苏省南通市通州区检察院搭建了“异常行为分析”云平台，其主要是对监管场所的异常行为进行自动预警。

［3］2022年11月1日，中国首部人工智能产业专项立法《深圳经济特区人工智能产业促进条例》开始实施。其措施上强调了设立市人工智能伦理委员会对人工智能领域的安全伦理问题进行监督。在标准规范上，强调了建立和完善政府规范、行业自律、企业自治、社会监督的人工智能治理机制，推动形成具有广泛共识的人工智能治理框架和标准规范。2023年国家网信办等七部门联合公布的《生成式人工智能服务管理暂行办法》，该办法主要从“网络安全”“数据安全”“个人信息”“科学进步”四个方面对生成式人工智能进行限制。2021年国家新一代人工智能治理专业委员会发布的《新一代人工智能伦理规范》，该规范提出了增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养6项基本伦理要求。针对自动化行政处罚的相关规定，2022年，国务院发布的《关于进一步规范行政裁量权基准制定和管理工作的意见》强调了要建立健全行政裁量权基准制度，严格规范行政裁量权基准指定权限。

［4］江南公安和南京市公安搭建了行政案件的自动量罚方法系统，按照统一的标准，结合裁量规则，自动生成行政处罚结果。

［5］左卫民．中国计算法学的未来：审思与前瞻［J］．清华法学，2022（3）．

［6］算法歧视是由于学者担心算法设计会被资本裹挟，数据来源中存在低质量数据，运行规则存在不合理，导致算法存在对嫌疑人物化的风险。

［7］算法黑箱是由于算法设计者难以解释算法的运行过程，算法设计公司不会对算法内容进行公开，执法者对于结果产出的原因无法解释。

［8］参见张博雯．人工智能辅助刑事诉讼决策的正当性及风险消解——以量刑辅助系统为视角［J］．北京警察学院学报，2024（6）．

［9］参见李学尧，刘庄．算法时代的法治路径：计算法学的规范性探索［J］．交大法学，2025（1）．

拨开人工智能法学的神秘面纱，实现“有则改之，无则加勉”的目标。

二、治安处罚裁量辅助系统之技术机理

知其然更需知其所以然。为深入剖析上述理论问题，本文立足人机协同视角分析问题。首先，明晰研究对象本身，通过技术发展的阶段划分，定位治安处罚裁量辅助系统的技术时代背景；其次，秉持马克思主义矛盾观中“从一般到特殊”的角度认识问题，剖析大模型的理论构建；最后，遵循“从特殊到一般”的解决思路，即当今阶段治安处罚裁量辅助系统的技术内核。

（一）治安处罚裁量辅助系统发展阶段

治安处罚裁量辅助系统不是凭空产生的，其应用，而是经历了四个发展阶段：第一代以法律课本为核心，公安在进行办案的过程中通过翻阅法条明确处罚依据；第二代是信息化应用阶段，公安机关将法律规范录入警务系统，建立相应的法律知识库，办案人员通过网络进行查询；第三代为大数据应用阶段，随着办案人员录入信息的增加，从一开始录入一个案件编号、案件详情，到现在涉案人员信息、涉案财物信息、证据材料信息等，通过建立专家系统，可以快速地搜索出来与将要处罚案件相关的法律规定及同类案例的处罚结果；第四代是大模型应用阶段，即将法律数据库和案例数据库投喂给机器进行学习，在不同架构的基础上，运用不同的算法，使大模型按照公安的处罚逻辑进行分析，最终进行概率输出。目前最新技术应用已处于第四阶段。

从技术发展逻辑来看，其大致历经三大阶段：第一阶段的技术逻辑是将违法事实编辑为代码输入系统，遵循“if-then”规则，建立专家系统，公安人员将违法事实输入进系统内后，系统会对应的输出结果[1]；第二阶段的技术逻辑是将违法事实拆解为“行为”和“后果”两大维度，再在两类维度下拆分为若干子行为、子后果，在对相关要素进行量化的基础上，建立该二维数组的基本函数计算公式[2]；第三阶段的技术逻辑是将深度学习，在投喂给机器数据后，机器将在特定架构和算法支持下，总结执法人员的处罚规律，再根据规律进行概率输出。[3]

（二）治安处罚裁量辅助系统大模型构建理论

治安处罚裁量辅助系统主要依靠大模型技术，而大模型的基础是神经网络（又称深度学习）。为降低参数运算量，深度学习算法需要依靠更加高效的技术架构，如循环神经网络（RNN）、卷积神经网络（CNN）、Transformer架构等，而目前的大模型主要依靠的是Transformer架构。所以在此基础上，先进行模型的预训练（无监督学习），即使用海量无标准的数据，让模型学习数据的通用特征和潜在规律；然后在预训练模型的基础上，使用针对特定任务的少量标注数据对模型进行进一步的微调（监督学习），使其适应具体任务需求。[4]训练数据的过程中离不开工程技术和基础设施的支撑：工程技术涵盖分布式训练技术、优化技术、数据处理技术、部署与推理技术，这些技术能够提高训练的效率和稳定性；基础设施即算力设施（高性能云计算平台），其与工程技术的结合，能够为数据训练营造高效、稳定的环境。

公开数据显示，GPT-3模型的参数量达1750亿[5]，训练成本约460万美元。[6]因此，一家中小企业很难研发自己的大模型，通常情况下，中小企业是根据开源的大模型，然后进行微调，将其垂直应用在各个领域，这样能够减少很多生产成本。

---

［1］参见何扬阳．自动化行政的裁量困境与解决路径［J］．数据治理与计算法学，2024（3）．

［2］参见周佑勇．裁量基准的技术构造［J］．中外法学，2014（5）．

［3］参见江国华，宋雅婷．论人工智能嵌入行政处罚裁量的逻辑与进路［J］．时代法学，2024（6）．

［4］参见刘光宇．生成式人工智能赋能行政裁量研究：构造、风险与策略［J］．南海法学，2023（6）．

［5］Open AI．Language Models are Few-Shot Learners［EB/OL］．（2025-05-03）．https://arxiv.org/abs/2005.14165.

［6］Lambda Labs．The Cost of Training GPT-3［EB/OL］．（2025-05-03）．https://lambdalabs.com/blog/demystifying-gpt-3.

（三）治安处罚裁量辅助系统技术逻辑

武汉迪赛威智能科技有限公司是在开源大模型的基础上进行微调的，垂直应用于公安领域。治安裁量辅助系统具体包括三个功能“智能组卷”“案卷质检”“处罚裁量”。具体操作流程如下：公安人员接到报案后，第一时间出警开展勘察、检查、询问等，当所有材料充分后，警察会在公安系统上进行备份。此时，该治安处罚裁量辅助系统会自动整合材料形成卷宗，然后对卷宗材料进行质量检查，最后对合格卷宗材料进行“基本事实”“基本情节”“其他情节”等进行系统自动提取与判定；在机器判定的基础上，系统依据相关法律规定与相似案例，自动输出“处罚起点”“处罚结果”“处罚意见”；最后经公安分局法制部门审核审批，审核通过后，系统自动生成处罚决定书。

针对上述内容，系统采用“数据层—算法层—应用层”三层架构：数据层负责整合公安系统备份的多源异构数据，算法层以深度学习算法为核心，应用层则实现三大核心功能。

1. 智能组卷模块

一是采用OCR识别技术处理扫描文档，通过视频关键帧提取算法解析执法记录仪或监控视频影像，结合BERT模型完成文本语义结构化；二是通过定义“当事人—行为—证据”三元组关系模型，构建法律知识本体[1]；三是基于Bi-LSTM+Attention机制实现案情要素自动匹配，生成符合《公安机关办理行政案件程序规定》的标准化电子卷宗。[2]

2. 案卷质检模块

第一是通过建立检查清单（Checklist）[3]，采用规则引擎（Drools）进行强制验证[4]；第二是基于知识图谱（Neo4j）[5]构建证据链推理网络，通过图遍历算法识别矛盾点[6]；第三是通过孤立森林算法检测非常规操作模式。[7]

3. 处罚裁量模块

第一步是情节要素抽取，目的是从案件材料中自动识别裁量所需的关键要素。建立法律文本语义模型，选择法律领域预训练模型对案件文本进行编码，捕捉法律术语的语义特征，然后在标注的法律实体数据集上微调，使其适配“证据链完整”等具体要素的识别任务。通过Bi-LSTM+CRF模型识别案件中实体与实体之间的关系，构建法律实体标注体系。[8]第二步是法律适用推理，目的是将提取的案情要素与法律条款匹配，确定处罚的依据。其技术逻辑是通过图神经网络来实现对法律知识图谱的学习，实现“违法行为—处罚条款—情节等级—裁量标准”的推理。[9]第三步是案例相似度计算，目的是检索历史相似案例，为裁量提供参考依据。技术逻辑包括案例表示学习和动态权重分配，这样通过更好地契合情节要素的差异化特征，优化相似度计算。[10]第四步是多目标裁量建议生成，目的是结合法律法规和相似案例等因素，生成处罚建议。技术逻辑是在设计法律符合度、案例一致性、社会效果评估三个目标函数的基础上，强化学习调参，最终输出合

---

［1］参见王治政，王雷，李帅驰，等．基于多视角知识图谱嵌入的量刑预测［J］．模式识别与人工智能，2021（7）：655-665.

［2］See Chalkidis I，et al．LEGAL-BERT：The muppets straight out of law school［J］．arXiv，2010：2559.

［3］参见王强，刘洋．基于规则引擎的政务数据质量管控系统设计［J］．计算机工程与应用，2019（12）.

［4］See Drools Team．Drools Rule Engine Overview［EB/OL］．（2023）．Drools Documentation.

［5］See Hogan A，et al．Knowledge Graphs［J］．ACM Computing Surveys，2021，54：1-37.

［6］参见李明，张磊．基于知识图谱的刑事证据链一致性校验方法［J］．中国司法鉴定，2020（5）.

［7］See Chen Y，et al．Anomaly Detection in Law Enforcement Data：A Case Study of Police Patrol Records［J］．IEEE Access，2021，9：123456-123467.

［8］参见李明，等．基于Bi-LSTM-CRF的行政处罚案件要素抽取方法［J］．计算机应用研究，2022（3）.

［9］参见王磊，等．基于GAT的多跳法律推理在行政处罚裁量中的应用［J］．人工智能学报，2023（2）.

［10］参见张伟，等．基于注意力机制的法律案例动态权重匹配模型［J］．计算机研究与发展，2023（5）.

理的处罚区间。[1]

三、技术预测辅助公安执法决策之风险检视

全方位对治安处罚裁量辅助系统定位后，本文从人机协同的视角，将学者们关于技术预测辅助法律决策的矛盾观点进行归纳总结。从决策推理的底层逻辑出发，衍生出相关性与因果性的矛盾，具体表现为大模型的推理逻辑是对相关性的提炼和预测，这与法律工作者的因果推论存在本质的不同，相关性一定程度上会引发“算法黑箱”问题。从决策的结果出发，衍生出客观性与主观性的矛盾，具体表现为大模型预测的结果相较于人类决策会更加客观准确。而完全基于文本的客观法律决策，可能会由于无法兼顾法律效果与社会效果的统一，引发实体风险。在此基础上，由于技术的便利性和高效性，衍生出独立性与依赖性的矛盾，具体表现为执法人员过度依赖技术引起的机械裁量和权力架空等程序风险。基于上述的程序、实体、算法风险，最终会导致潜在的法律风险即归责真空问题。综上，本文在大模型的技术背景和人机协同的互动视角下，对上述风险进行再次审视与思考，为治安处罚裁量辅助系统进行更加精准的风险定位。

（一）技术预测辅助法律决策的矛盾剖析

1. 相关性与因果性

相关性不等于因果关系，如果错误地推断因果关系，可能会造成严重的后果。[2]机器学习的法律计算能从纷繁复杂的历史数据中发现并总结规律，当有相似的案件出现时，会给出相近的结果，本质上是发现数据之间的相关性，而非揭示因果性。

目前法律大语言模型存在以下应用场景，如法律检索、类案推送、文书撰写、案件说理等。[3]上述需求可以分为两大部分：一部分为“搜索引擎”，另一部分为“思考判断”。前者通过文本分析判断相关程度，进而进行推送，后者体现出大模型应具有法律知识，能够进行思考并给出相应的答案。但大模型本质上是一种预测模型，与前者的诉求存在本质区别。通过大量的文本训练，预测下一个字或句子的概率分布。基于此，大模型会出现形式上“头头是道”、实质上“胡编乱造”的情况，给法律工作者的实践应用带来很大的考验。

2. 客观性与主观性

机器预测结果可能相较于人类决策会更加客观。机器预测是指对以往的案件、法规等数据通过决策树、梯度提升算法等机器学习方法，构建和训练计算模型，进而实现对法律结果的预测。相较于机器预测，人类决策会考虑“不具有观察性”的要素，进而产生结果上的偏差。以法庭审理为例，法官会在一定程度上受到潜在因素影响，如被告在法庭中的行为等，进而产生系统性偏差。例如，机器学习的预测结果可能有助于减少监禁率与犯罪率。[4]

3. 独立性与依赖性

机器预测具有使用便利性、运行高效性、结果客观性的特点，这极大地激发了执法人员使用技术的积极性，但过度依赖机器决策会出现机械裁量的情况。在人工智能应用成为主流趋势的背景下，随着量刑辅助系统的不断应用，执法人为了追求“同案同判”“高效办案”，会增加对系统的依赖程度，这在一定程度上削弱执法人员的主体地位。同时，机器执法的底层逻辑是代码，学习的文本素材是以往相似的裁判文书，兼具权威性与高效性，会较为容易地引发机械裁量现象。[5]

（二）技术预测辅助法律决策的风险检视

1. 程序风险之辨

关于程序正义问题，大多数学者认为，自动化行政处罚会将原来的传统执法模式变为线上执法模式，行政人员会成为“机器判官”。由于执法人员无法到现场，行政相对人无法为自己进行辩护，导

---

［1］参见刘强，等．基于多目标优化的行政处罚智能裁量模型研究［J］．计算机应用研究，2022（8）．

［2］参见刘庄．幻象与本相：法律人工智能及其他［J］．中国法学评论，2024（2）．

［3］参见刘庄．从“世界模型”看人工智能在法律场景的实际应用［J］．中国应用法学，2024（2）．

［4］See Kleinberg J，Lakkaraju H．Human Decisions and Machine Predictions［J］．The Quarterly Journal of Economics，2018，133（1）：237–293.

［5］参见张会平，曹景伟．算法行政对自由裁量权的影响：正向规制、负向限缩与动态平衡［J］．信息技术与管理应用，2023（3）．

致自己的合法辩护权受到了侵害。

根据《治安管理处罚法》，公安机关在办案的过程中需要遵循“受案—调查—决定—执行—执法监督”的程序。其中第九十四条规定，公安机关有告知违反治安管理行为人作出治安管理处罚的事实、理由及依据的义务，违反治安管理行为人享有陈述和申辩的权利。根据《公安机关办理行政案件程序规定》第一百六十七条、第一百六十八条、第一百六十九条规定，公安机关具有告知义务，违法嫌疑人具有陈述和申辩的权利。第一百七十条和第一百七十一条规定公安法制部门应当对行政案件进行审核。

本研究在调研的过程中发现，治安处罚裁量辅助系统的应用场景为受案调查结束后、行政处罚决定前。具体来说，该治安处罚裁量辅助系统只承担“智能组卷”“卷宗质检”“处罚裁量”三个部分的任务。最终，公安分局的法制部门对行政案件进行审核与审批，审核通过后，系统自动生成处罚决定书。

调研的过程显示，该系统的应用保留了传统办案的优势，即和行政相对人交互的现场感，又通过大模型减少了归类卷宗、卷宗质检、查阅法条与案例的时间，提高了办案效率。所以，没有侵犯或剥夺违反治安管理行为人的辩护权，反而在此基础上实现了办案效率的优化。

2. 实体风险之辨

目前，很多学者认为裁量辅助系统是通过识别法律代码来进行处罚裁量工作的。但法律语言属于自然语言，具有不确定性，而机器语言具有确定性，不允许二义性的存在。[1]所以在实践中，学者们担心法律代码化无法实时更新导致错判。如“姜国富与攀枝花市公安局交通警察支队一大队公安道路行政处罚纠纷上诉案”。[2]学者们担心法律编码错误导致错判，如“嘉峪关浩峰能源有限责任公司与嘉峪关市环境保护局行政处罚案”[3]和“肃北县凯富矿业有限责任公司与酒泉市生态环境局肃北分局案”[4]。针对上述问题，学者们的讨论还局限于“专家系统”的辅助审判阶段。该阶段对法律代码化的要求极高，一旦出现翻译错误或翻译不完全，就会出现错判的现象。所以，当前随着技术的不断发展，法律代码化已经从传统的对照转译型、细化转译型，转变为创设转译型[5]，即通过深度学习的算法模型，自主判断法律行为的违法性和相应的法律后果。

在深度学习算法模型的讨论框架下，治安处罚裁量辅助系统能否保证结果输出的正确性呢？理论上，由于深度学习本质上是对大量数据的非线性拟合，在此基础上进行多层网络逐级抽象特征提取与学习，所以无法做到完全准确。在数据层面上，根据大数定律，当样本量$n\to\infty$时，样本均值收敛于总体均值。在最小二乘法中，随着数据量增加，估计值趋近于真实值。所以在深度学习中，可以通过大量样本逼近真实数据分布，降低模型方差，提高泛化能力。在数据层面上，随着机器的持续学习与更新，能够提高结果的准确性。在模型层面上，可以通过模型结构调整、训练策略优化、分布式训练优化逐步提高模型训练的参数量并降低训练成本。在技术工程层面上，可以通过参数高效微调（PEFT）和提示工程进行强化学习。

综上所述，深度学习算法相较于传统的转译算法而言，会通过不断学习，增加结果输出的灵活性与准确性。

3. 算法风险之辨

人工智能本质上是针对不同情景给出针对性的输出反应，其载体就是使$f(x)=y$成立的黑箱。换言之，在庞大的数据输入后，人工智能进行数据分析，最后进行决策，而其中人工智能学习的过程因为无法透明化，而被称为“算法黑箱”。同时，由于数据输入具有质量不高的可能性，数据中隐含着结构性歧视，如种族、性别、年龄歧视等，所以输出的结果也不可避免地具有“歧视色彩”，这个过

[1] 参见刘东亮．技术性正当程序：人工智能时代程序法和算法的双重变奏［J］．比较法研究，2020（5）．

[2] 参见四川省攀枝花市中级人民法院（2016）川04行终13号行政判决书。

[3] 参见甘肃省酒泉市中级人民法院（2018）甘09行初2号行政判决书。

[4] 参见甘肃省敦煌市人民法院（2019）甘0982行初23号行政判决书。

[5] 参见魏文杰，余洋．数字化时代行政裁量治理：价值、隐忧与法律规制［J］．江西师范大学学报（哲学社会科学版），2024（3）．

程称为“算法歧视”。

针对“算法黑箱”的问题，在机器学习的过程中确实存在着难以解释的现象，如为什么会出现这样的结果，机器是怎么进行学习的？但是，法官或警察等执法人员在执法的过程中，作为第三者也很难得知执法人员的思考过程，也存在着所谓的“思考黑箱”。但当事人不担心该黑箱，原因在于执法人员会通过详细的论证来证明处罚的合理性。所以，只要能够通过各种手段让思考过程可视化，论证行为与结果之间的逻辑关系，“黑箱”并不可怕。而目前的治安处罚裁量辅助系统能够通过文本分析，联系相应的法规与案例，给出论证过程，最后得出处罚结论。因此，在这样的视角下，“黑箱”的问题能够得到有效解决。

针对“算法歧视”的问题，需要区分清楚“结构化歧视”和“人际歧视”。以美国的COMPAS系统为例，美国是资本主义国家，国家的权力和财富掌握在资本家的手中。虽然强调“天赋人权”与“人人平等”，但是由于历史因素，一直存在着种族歧视现象。而这种“种族”“肤色”“国籍”为结构化歧视，COMPAS系统将其纳入参数设计，所以其结果存在着较大的非正义性，如损害公民的平等权利。但是我国是人民民主专政的社会主义国家，同时宪法中规定国家尊重和保障人权，不存在所谓的“结构化歧视”。在司法领域，宪法要求公正司法。所以在数据层面，我国的案例数据质量较高。但我国国司法实践具有“审判中心主义”的特征，案例具有一定程度的“个人色彩”，即“人际歧视”。但是随着样本数据量的增加，大数法则定理下，样本均值收敛于总体样本均值，会减少“个人色彩”的影响，达到客观的效果。

4. 人机协同引发的责任真空

在法律层面，国内对于人工智能的规定仅仅停留于宏观制度的把控，没有在具体应用领域进行细化规定。在自动化行政处罚领域，国家在宏观上对方向进行了把控，即建立健全行政裁量权基准制度，严格规范行政裁量权基准指定权限。[1]但在实践层面却未对智能辅助系统的法律地位进行明确。如现行的《行政处罚法》《治安管理处罚法》《公安机关办理行政案件程序规定》等法律法规，对“治安处罚裁量辅助系统”的法律属性没有明确。这会引发两类主体的归责风险，针对警察来讲，目前法律法规对其责任有所规定。[2]但对于第三方主体，如第三方审计、系统缺陷设计者，出现了归责漏洞。综上，目前法律法规尚未对AI辅助决策的责任分配规则作出明确的规定，存在着责任归因上的法律真空。

四、治安处罚裁量辅助系统之风险应对策略

人工智能的发展不会因为某些不可突破的桎梏而停滞，只会通过某些策略而规制。人机互动的趋势是不可避免的，因此人机协同理论显得尤为重要。人机协同理论强调人与机器共存，既要发挥人工智能技术的快速数据处理优势，又要发挥公安执法灵活性的优势，实现优势互补。目前，随着半自动化行政裁量向全自动化行政裁量的发展，人与机器的地位发展也经历了三个阶段：第一阶段是“人在机器决策中”的形式，即机器在计算处罚结果的过程中需要人为设定参数、输入内容；第二阶段是“人在机器决策上”的形式，即机器完成决策的所有过程，人不参与其中，只进行监督与控制；第三阶段是“人在机器决策外”的形式，即人不干预机器出发的结果，只进行前期的投喂数据与模型训练。[3]目前，从大模型应用阶段来看，其仍处在“人在机器决策上”的阶段，但是可以学习“人在机器决策外”的模式。

基于上述对治安处罚裁量辅助系统的定位，以及前文分析的潜在风险，本文通过人机协同策略来进行风险应对策略：从“人”的角度具体探析场景应用策略，从“机”的角度具体明确算法适用策略，最后从制度层面对“人”与“机”进行程序监督。

---

［1］参见《国务院办公厅关于进一步规范行政裁量权基准制定和管理工作的意见》国办发〔2022〕27号。

［2］参见《公安机关人民警察执法过错责任追究规定》（2016年1月14日公安部令第138号发布 自2016年3月1日起施行）第五条：“执法办案人、鉴定人、审核人、审批人都有故意或者过失造成执法过错的，应当根据各自对执法过错所起的作用，分别承担责任。”

［3］参见陈悦．人机协同裁量的构建原理风险检视与因应优化［J］．浙江学刊，2024（5）．

（一）场景应用策略

第一，适用于基层公安执法场景。基层公安存在着“案多人少”的现象，警力不足成为公安办案过程中的主要难题。所以，在充分发挥警员工作能力的同时，提高办案的效率与质量成了当前的重中之重。治安处罚裁量辅助系统能够很好地解决这个问题，不仅能够帮助见习民警快速上手，而且能够快速地整合卷宗与质检，提高了办理案件的效率。

第二，适用于简单且重复性强的治安案件。基层派出所办理的案件中，有很大一部分是打架斗殴、赌博等简单且重复发生的案件。治安处罚裁量辅助系统能够高效又准确地处理这类案件，释放警力资源，更好地让警力集中到复杂疑难的案件中。

（二）算法适用策略

针对法律动态更新的挑战，可以通过设计增量学习机制，以实现法律知识图谱的实时同步与冲突消解。法律知识图谱的增量学习通过GCN（图卷积神经网络）识别新增的法律节点范围，然后通过逻辑一致性验证，检查新条款与现有规则是否存在冲突。

针对规避“机械裁量”的挑战，可以构建处罚效果长期追踪与多维度反馈机制。该机制通过数据采集、融合建模、反馈优化的闭环架构，实现处罚效果的动态评估与模型迭代。从数据采集的角度，可以增加数据维度，不仅包括法律库和案例库，还可以增加反映社会效果的新闻评论数据等。从构建动态效果评估模型的角度，可以构建损失函数评估效果模型达到处罚效果满足多维平衡。例如，$L=\alpha$ *法律合规损失+β *社会风险损失+γ *经济成本损失，法律合规损失即处罚决定被行政复议的比例，社会风险损失即舆情负面情感指数和发生的概率，经济成本损失即警力投入的执行成本和相对人的经济损失。从反馈闭环与模型优化的角度，可以通过区块链技术记录每次处罚结果及其效果数据，确保反馈链路不可篡改。在此基础上，通过强化学习民警的反馈数据和高频率的人工修改记录，动态调整模型参数。

（三）程序监督策略

针对算法黑箱的挑战，可以构建“算法透明强制披露—第三方鉴定—立法标准化”三位一体的程序监督策略。首先，对于对处罚结果输出影响较为核心的算法，技术公司应向法院或第三方司法鉴定机构提供API接口进行验证；其次，由具有鉴定资格的人通过对抗性测试，检验大模型的置信区间等，然后给出司法鉴定报告；最后，《治安管理处罚法》中可以嵌入“数据质量审查”“算法可解释标准”“算法安全分级”等特别程序条款。

针对责任归因真空的挑战，可以构建“决策留痕—角色锚定—责任分层”三位一体的程序监督策略。首先，可以将治安处罚裁量辅助系统的决策过程按照电子数据真实性的标准进行全程留痕[1]；其次，通过《公安机关办理行政案件程序规定》规定开发者定期进行算法披露和提供算法评估报告的义务，公安法制人员对于结果的审核义务，第三方鉴定机构对于算法披露的鉴定义务；最后，在上述方式的基础上，若存在追责情况，不仅可以查明哪个环节出现了问题，还可以将责任落实到对应的主体。

五、结语

本文以人机协同为视角，以“承认应用技术与发挥人的主观能动性并举”为行文前提，以马克思主义矛盾观构成行文逻辑。首先从一般到特殊，本文先明晰治安处罚裁量辅助系统的所处阶段为大模型阶段，所应用技术为深度学习技术，进而讨论大模型的构建理论，即基于Transformer架构的模型预训练与微调，其次从特殊到一般，具体讨论治安处罚裁量辅助系统的技术逻辑，通过“数据层—算法层—应用层”三层架构，剖析“智能组卷”“案件质检”“处罚裁量”三个功能。在认识该技术的机理后，本文将理论界对于人机协同模式的质疑进行观点归纳，大致分为相关性与因果性的矛盾、主观性与客观性的矛盾、独立性与依赖性的矛盾。以此为逻辑原点，本文推导出理论界的三大风险“共识”，即程序风险、实体风险、算法风险。针对上

[1] 参见最高人民法院．关于互联网法院审理案件若干问题的规定（2018）[EB/OL]．法释[2018]16号第十一条：当事人提交的电子证据，通过电子签名、可信时间戳、哈希值校验、区块链等证据收集、固定和防篡改的技术手段或者通过电子取证存证平台认证，能够证明其真实性的，互联网法院应当确认。

述风险，本文基于新技术背景与人机协同视角对上述风险进行再审视与思考，得出以下结论：程序上没有侵犯或剥夺违反治安管理行为人的辩护权，反而进行了办案效率的优化；实体上深度学习算法相较于传统的转译算法而言，会通过不断学习，增加结果输出的灵活性与准确性；算法上“黑箱”可以通过文本分析，联系相应的法规与案例，给出论证过程及逆行有效地缓解。“歧视”基于我国国情和司法体系，具有“人际歧视”的特点，其可以通过大数法则减少“个人色彩”的影响，以达到客观的效果。与此同时，本文通过检视后提出了一种责任真空的法律风险，即目前法律法规尚未对AI辅助决策的责任分配规则作出明确的规定，存在着责任归因上的法律空白。

综上对风险的分析，本文基于人机协同理论，在风险应对上将人与机器的关系分为三个阶段。目前，从大模型应用阶段来看，其仍处在“人在机器决策上”的阶段，但是可以学习“人在机器决策外”的模式。具体表现为：从“人”的角度探析场景应用策略，执法人员可以在基层公安执法场景和简单且重复性的治安案件中使用；从“机”的角度明确算法适用策略，通过设计增量学习机制来应对法律动态更新的挑战，可以构建处罚效果长期追踪与多维度反馈机制规避“机械裁量”的挑战；从制度层面对“人”与“机”进行程序监督，可以构建“算法透明强制披露—第三方鉴定—立法标准化”三位一体的程序监督策略应对“算法黑箱”的挑战，以及构建“决策留痕—角色锚定—责任分层”三位一体的程序监督策略应对责任归因真空的挑战。总之，共同发挥二者的作用，在治安处罚裁量辅助系统辅助全自动化行政处罚的基础上，既要发挥公安分局法制部门的监督作用，又要发挥技术人员在前期数据预训练与微调的作用，这样不仅可以提高机器的学习能力，也能够保证输出结果的准确性。

时代的车轮滚滚向前，科技如同发动机为社会发展提供动力，那么驶向何方就是本文讨论的重点。国家强调要加强人工智能与执法活动的紧密结合，而当今与众不同之处在于，技术革命如火如荼，生成式人工智能的应用重塑了技术辅助执法的底层逻辑，即由之前“专家主义”转向现在的“连接主义”。这也为重新审视大模型在各领域的应用风险提供了契机。各领域亦有不同：相较于政府其他部门的客观程序，公安在决策时需要兼顾主观能动性与客观程序；相较于法检系统技术应用的成熟状态，公安刚刚进行技术转型。这可能是由于法检作为司法部门，办理案件时具有被动性的特点，可以在案件事实明确后进行高效办案，而公安作为社会治安维护者，具有主动预防与执法的特点。所以这也体现出大模型技术在公安领域的应用，具有其他主体不可比拟的特殊性。因此明晰治安处罚裁量辅助系统的技术原理，思考潜在风险及应对相应风险策略，是值得反复推敲的课题。

（责任编辑：李秀玲）

Technical Principles, Risk Review and Response Strategies of Public Security Punishment Discretionary Assistance System from the Perspective of Human-Machine Collaboration

Chang Keqiang

School of Criminal Justice, Zhongnan University of Economics and Law, Wuhan

Abstract: The integration of artificial intelligence technology and public security law enforcement has become an inevitable trend. Through the field research on Wuhan DearSeeWe Intelligent Technology Co., Ltd, this paper finds that current public security law enforcement is evolving from the "substantive decision-making" stage to the "substantive prediction" stage, and the technical application is also transforming from "expertism" to "connectivism". However, most existing academic research focuses on the legal and procuratorial systems and "expertism"-related fields, and research on the application of "connectivism" technology in public security law enforcement is still blank. Taking this as the entry point, this paper first clarifies the technical principles and construction connotation of the public security punishment discretionary assistance system; second, taking correlation and causality, subjectivity and objectivity, independence and dependence as the core contradiction origins, it summarizes three major risk types recognized in the theoretical circle: procedural risks, substantive risks and algorithmic risks; finally, combined with the new application background of legal large models, it re-examines the above risks, and supported by the "human-machine collaboration" theory, puts forward corresponding risk response paths from three dimensions: application strategy, algorithmic strategy and procedural strategy, aiming to realize the organic combination of technical advantages and the professional advantages of law enforcement personnel and achieve a win-win situation.

Key words: Human-machine collaboration theory; Computational law; Technical principle; Risk review; Response strategy