

犯罪治理算法的法律规制

单 勇^{*}

内容提要 犯罪之算法治理已成为平安中国建设的基本方略,其应用覆盖地点评分、维稳风控、大数据扫黄、网络犯罪风控等场景。以“监视—筛选”为特质的犯罪治理算法在极大提升治理能力的同时,也存在法律依据不健全、对个体生活形成挤压和算法权力失范等问题。为推动算法治理从“基于技术的治理”转向“基于规则的治理”,应引入以结果公平和过程公平为内涵的算法公平原则,作为建构“算法之法”的正当性基础。依循回应性规制和预防性规制路径,通过算法救济和问责等回应性机制修复算法偏误对结果公平的侵蚀,依托算法影响评估和算法公开等预防性机制防止算法失范对过程公平的偏离,以促进算法向善、保障算法治理始终运行于法治轨道。

关键词 犯罪治理算法 算法治理 算法公平 回应性规制 预防性规制

DOI:10.14111/j.cnki.zgfx.2026.02.004

目 次

- 一、犯罪治理算法的实践检视
- 二、源自算法公平的正当性思考
- 三、犯罪治理算法的回应性规制
- 四、犯罪治理算法的预防性规制
- 五、结语

^{*} 南京大学法学院教授。本文系 2025 年度江苏省社科基金重点项目“人工智能嵌入犯罪治理的法治化研究”(项目批准号: 25ZHA008) 的阶段性成果。

一、犯罪治理算法的实践检视

随着“算法渗透型社会”^①（algorithmically infused societies）的到来，运用算法技术预防和控制违法犯罪的算法治理成为平安中国建设的新兴模式，更构成了人工智能时代犯罪治理体系的底层逻辑。算法既是“人类通过代码设置、数据运算与机器自动化判断进行决策的一套机制”^②，也是人工智能系统的核心技术，还是“塑造现实世界的‘隐形建筑师’”^③以及犯罪治理的通用性、基础性工具。如今，算法技术已深嵌于社会治安整体防控体系，犯罪治理算法的泛在应用催生出“算法决策+人工干预”的智治模式。

（一）犯罪治理算法的应用图景

犯罪治理算法的实践依循“数据—算法—平台—干预”流程，以“在线控制”引导对社会风险和违法犯罪的“在场控制”。基于对“人、地、事、物、单位、行业、网络”等要素的数据解析，个体被冠以数字身份而成为算法解析对象。以过滤类算法为内核的判别式人工智能系统可以自动化处理数据，并基于监测、比对、判别、筛选、分类、排序等功能精准识别违法犯罪风险，通过风险评分与风险叠加等方式输出结论，再结合人工判断，完成从数据到决策的闭环。算法或基于政法机关的治理平台运行，或依托承担“看门人义务”^④的互联网平台落地，这两类平台由此成为算法治理的信息中枢、组织枢纽，其中，国家机关与互联网企业系常见的算法部署者、使用者。目前，犯罪治理算法广泛用于街面犯罪地点评分、大数据反腐、维稳风控、大数据扫黄和网络犯罪风控等场景。其中，地点评分较少使用个人信息，而其他应用则多以个人信息作为风险研判的依据。

第一，街面犯罪的地点评分。算法治理的最初探索是对盗窃、抢劫等街面犯罪进行空间分析，以地理信息系统分析犯罪空间分布特征，制作犯罪地图、探测犯罪热点，运用空间回归等模型揭示“城市吸引犯罪”的内在规律。^⑤例如，“美国加利福尼亚州圣克鲁斯县使用 PredPol 系统，通过分析历史数据将警方指向 10—20 个高风险犯罪区域。在该系统使用的第一年就减少 27% 的入室盗窃案件和 11% 的住宅入室盗窃案件。”^⑥我国公安机关也广泛运用地点评分，评估城市空间犯罪风险、优化警力部署、改善防卫空间设计。

第二，大数据反腐的算法风控。大数据反腐是数字纪检监察的通俗表达，以公权力运行数据轨迹和个人异常风险为研判依据，依循“数据—模型—线索—案件—治理”的逻辑，实现对公权力运行的在线监督、主动监督、穿透式监督、全链条监督。一方面，以算

① Claudia Wagner et al., *Measuring Algorithmically Infused Societies*, *Nature*, Vol.595:197, p.197 (2021).

② 丁晓东：《论算法的法律规制》，载《中国社会科学》2020年第12期，第141页。

③ 张欣：《算法公平的类型构建与制度实现》，载《中外法学》2024年第4期，第876页。

④ 单勇：《论网络犯罪的看门人规制》，载《南京大学学报（哲学·人文科学·社会科学）》2023年第5期，第62页。

⑤ 参见单勇：《城市高密度区域的犯罪吸引机制》，载《法学研究》2018年第3期，第119-120页。

⑥ Federico Liberatore, Miguel Camacho-Collados & Lara Quijano-Sánchez, *Towards Social Fairness in Smart Policing: Leveraging Territorial, Racial, and Workload Fairness in the Police Districting Problem*, *Socio-Economic Planning Sciences*, Vol.87:1, p.1 (2023).

法技术解析公权力运行数据轨迹。S市纪委监委深化粮食购销领域监督,梳理行权流程,采集3亿条数据,按“人、企、车、粮、钱”划分,建立17个监督模型,发现“转圈粮”“损耗粮”等线索105件,自查问题220个。^⑦另一方面,依托算法识别干部个人异常风险。H市纪委监委“党风廉政建设和反腐败信息共享平台”涵盖上千名市管干部、47万余名党员干部及各级公务员的基本信息,对每名市管干部采集个人信息达256条,延伸链接房产、税务、水电气缴费等多部门数据,通过大数据排查,为执纪审查提供信息支持和决策参考。^⑧

第三,重点人员的维稳风控。以算法助力对“三失一偏”^⑨人员的服务管理已成为化解矛盾风险、防范“民转刑”“刑转命”案件的关键机制。据调研获知,某市公安局依据警源、诉源、访源、政府部门流转等来源,打造重大矛盾纠纷和问题隐患全量数据库,研发预防极端案件的算法模型,以上千个风险标签组成标签话语体系,实现防范重大风险和极端案件的最小颗粒度划分。维稳风控通过“AI+人工判断”方式对矛盾风险自动化分拣、打标、识险与研判,以风险标签叠加为预警规则,形成对个案的三色评估结论(绿色是已化解低风险,黄色是已化解仍需关注,红色是风险未化解需重点关注),指导派出所和社区民警精准化解风险。2024年,该市重点矛盾纠纷的化解率为99.72%,故意伤害刑事案件下降18.15%,涉及民事纠纷或精神障碍患者肇事肇祸引发的命案下降18.18%,最大限度把风险化解在基层,将极端案件阻截在源头。

第四,大数据扫黄的预警评分。完整的犯罪治理包括风险识别和预防处置两个阶段,其中评分机制是对治理对象的解析、评级及排名,构成风险识别的关键环节。大数据扫黄中的预警评分机制颇为典型,某地公安机关以“旅馆业治安管理信息系统”为基础,通过联动其他信息系统进行相关性分析,对高危涉黄人员开展预警评分。常见的做法是将高危人员信息与宾馆入住信息进行交叉比对,监测娱乐场所有偿陪侍人员等的宾馆入住登记情况,通过对开房次数和时间、异性同住人员情况、有无涉黄违法记录等指标设定权重并进行评分累积,对触达预警值的卖淫嫖娼案件线索予以推送。

第五,网络犯罪的看门人在线风控。前述算法治理是国家机关对违法犯罪的直接控制。为应对网络犯罪挑战,“国家管平台、平台管用户”治理模式应运而生,国家通过互联网平台实施间接控制的模式日臻完善。“平台管用户”的依据源于《反电信网络诈骗法》等法律为平台设定的犯罪控制义务,实质在于互联网企业对用户违法犯罪的在线控制。平台以用户的操作流、信息流、资金流、管理流等在线活动为监测对象,运用风险画像、行为相似度算法、账号聚集度算法、工作室识别算法等,履行内容审核、实名核验、协助执法、黑灰产治理等义务。据调研获知,国内某生活服务平台聚焦全景画像、行为分析、关系网络、异常检测等领域,运用对用户、商户、骑手的地址和设备等数据的全景画像

⑦ 参见李灵娜:《3亿条数据让“蛀虫”现形》,载中央纪委国家监委网站, https://www.ccdi.gov.cn/yaowenn/202203/t20220323_179503_m.html, 2025年8月1日访问。

⑧ 参见郑少忠、柯仲甲:《反腐有了大数据“侦探”》,载《人民日报》2017年8月15日,第10版。

⑨ “三失一偏”人员指生活失意、心态失衡、行为失常、性格偏执人员。参见曹莉、刘莉:《南宁市破解“三失一偏”人员服务管理难题》,载《长安》2025年第6期,第45页。

能力,自动标识多维度、多类型特征标签,快速定位识别风险人员;以机器学习分析用户行为,实现线上行为感知、线下轨迹预测和聚集事件识别;基于图数据库和知识图谱,构建可推理、可量化、可解释的复杂关系网络,挖掘同地址、同设备、同网络等团伙关联关系和多层级实体关系;利用画像、行为、网络结构等数据,自动化监测网络黑灰产异常情况。由此,平台打造针对“杀猪盘”、刷单引流诈骗、洗钱、涉黄、涉毒等违法犯罪的风控模型,形成全流程安全风控体系。

(二) 犯罪治理算法的规制之问

犯罪治理算法通过监视社会及个体的数字轨迹,基于风险叠加或评分机制从高危、异常乃至“无害”活动中过滤违法犯罪风险,形塑出以“监视—筛选”为特质的算法权力运行逻辑。“在大数据技术和算法技术广泛使用的背景下,所有人的生物性生命都被数据化,成为基于数据库治理的控制社会,这是智能时代的数字—生命政治……”^⑩问题在于,个体被算法细致入微地记录分析,虽极大提升了国家洞察社会及个体生活的广度与深度,但社会治理不能以效率为单一导向,要警惕借数字利维坦躯壳复活全能政府的做法。“我们是否应该以及在多大程度上应该抵拒和限制技术治理和算法自动化决策对社会秩序的塑造……如何在技术治理所带来的收益与其对道德价值的可能减损之间作出权衡?”^⑪算法权力的扩张引发了关于权利保障的担忧和信任危机。一方面,算法控制的起点是“将人和物全面数字化并最大限度获取个人数据”^⑫,“对社会主体行为数据的全面收集和分析,可能会对相关主体自主生活空间形成挤压,影响到社会正常活动的进行”^⑬。犯罪的算法治理催生出“以大规模监控为基础的预防式执法,微观层面可能导致对公民私人空间的过度窥探,宏观上则可能带来社会的寒蝉效应”^⑭。另一方面,算法治理的功能偏重预警预防,多用于案件线索发现且处于刑事立案之前,面临相当程度的法律规制空白。某些算法实践既欠缺完备的法律依据,也存在“重算法创新”“轻权力制约和权利保障”的隐忧。算法治理在形塑高效智能的预防性秩序之时,也引发了其是否违背法治原则的争论。

党的二十届三中全会提出“全面推进国家各方面工作法治化”的要求,而推进犯罪治理现代化蕴含着法治化的当然之义。面对算法创新在先、法律相对迟滞的困境,“控制社会控制”^⑮的法治逻辑呈现“控制算法控制”的新要求。“在算法社会,规制的核心问题不是算法,而是使用算法的人及允许自己被算法支配的人。”^⑯在此意义上,无论是“规制算法治理”还是“规制犯罪治理算法”,本质上均强调对算法权力的规范和约束,算法治理必须从“基于技术的治理”转向“基于规则的治理”。以往研究对商业算法规

^⑩ 蓝江:《什么是生命政治》,载《武汉大学学报(哲学社会科学版)》2022年第1期,第57页。

^⑪ [英]罗杰·布朗斯沃德:《法律3.0:规则、规制和技术》,毛海栋译,北京大学出版社2023年版,代译序第9页。

^⑫ 王敏芝:《算法之下:“透明社会”的技术与观念》,载《探索与争鸣》2021年第3期,第29页。

^⑬ 马怀德:《数字法治政府的内涵特征、基本原则及建设路径》,载《华东政法大学学报》2024年第3期,第15页。

^⑭ 王锡锌:《数治与法治:数字行政的法治约束》,载《中国人民大学学报》2022年第6期,第22页。

^⑮ 白建军:《关系犯罪学》(第3版),中国人民大学出版社2014年版,第333页。

^⑯ [美]杰克·巴尔金:《算法社会中的三大法则》,刘颖、陈瑶瑶译,载《法治现代化研究》2021年第2期,第188页。

制关注颇多,对犯罪治理算法的规制尚处于起步阶段。学界注意到算法治理的“双刃剑”效应,呼唤约束算法权力和建构“算法规制体系”^{①⑦},强调“既要注重对国家等主体的权力制约,也要注重对平台等社会主体掌握的数据权力、人工智能技术的运用权力进行制约”^{①⑧}。许多学者也认识到算法应用的场景化特性,提出“算法规制必须采取场景化的规制路径”^{①⑨}的观点。由此,犯罪治理算法的法律规制成为一个不能被遮蔽的数字法治重点。为防范算法及人工智能引发的法治风险,国家必须以政策和法律主动塑造其发展轨迹。规制算法的目的不是否定算法创新,而在于通过审视算法治理的正当性,厘清算法权力与个人自由的边界,规范和改善算法治理,实现社会安全防卫与国民权利保障之均衡,使其发挥更重要作用并始终运行于法治轨道。

应当注意的是,犯罪治理算法在保护社会安全的过程中对个体权益保障造成了潜在风险,其与商业算法之间存在较大差异,对之规制是一个不易实现的难题,必须回答如下问题:犯罪治理算法存在何种法治风险?规制的原则、依据和方式有哪些?当犯罪治理算法出现偏误并产生负面影响时,如何开展救济和问责?在算法部署前及运行中,如何前瞻性识别潜在的权益保障风险?怎样在兼顾保密性要求的同时保障算法运行透明度?总之,通过法律规制实现犯罪治理算法的可用性(效益最大化)与可信性(风险最小化)的均衡,已成为推进犯罪治理现代化和法治化的核心问题。

二、源自算法公平的正当性思考

上述规制之问关涉到犯罪治理算法的数据可靠、风险可控、可解释、可审查、可救济、可问责、公众可参与等一揽子规制要求,指向对受影响者遭遇不公平对待的回应与对算法权力失范的预防,最终聚焦于“公平性”这一法律政治传统的根本价值。我国《全球人工智能治理倡议》《人工智能安全治理框架(2.0版)》和欧盟《人工智能法案》等政策法规均将“公平性”作为算法规制的基本原则。算法公平成为算法规制体系建构的理论原点和价值基座。在犯罪之算法治理领域,算法公平的法治意蕴尤为凸显。随着算法治理逐渐成为社会治安整体防控的底层行动逻辑,公众对算法可能过度干预个人生活的担忧与日俱增;但若对犯罪治理算法采取严苛限制立场,又影响对公共安全的拱卫,使国民暴露在违法犯罪风险之下,“当某些公平性措施显著损害公共安全时,这些措施本身亦非真正的公平”^{②⑩}。算法公平的核心要义在于,确保犯罪治理算法有序运行于预防性秩序建构和正当性权益保障的平衡之中。

(一) 犯罪治理算法的公平原则

关于算法公平的研讨,以数据正义为理论原点,荷兰蒂尔堡大学教授 Taylor 提出的

^{①⑦} 许可:《算法规制体系的中国建构与理论反思》,载《法律科学》2022年第1期,第124页。

^{①⑧} 杨建军:《可信人工智能发展与法律制度的构建》,载《东方法学》2024年第4期,第107页。

^{①⑨} 前注②,丁晓东文,第140页。

^{②⑩} Pak-Hang Wong, *Democratizing Algorithmic Fairness*, *Philosophy & Technology*, Vol.33:225, p.230 (2020).

“数据正义”三项原则构成了关于数据和算法使用之正义准则的基础性理论,即可见性、技术参与和反对不公平对待。可见性不仅涉及过度“可见”而引致的隐私风险,也关乎代表性问题,即保证数字弱势群体在数字系统中平等地被看见。技术参与强调个人在技术选择上的自主权,主张跨越数字鸿沟、共享数据红利。反对不公平对待原则要求在结果上避免偏误和歧视,还主张增强公众及监管机构识别、质疑、问责和纠正数据误用的能力。^②算法是数据处理的基本方法,算法治理亦是一种数据使用行为,故而数据正义三项原则为算法治理的法治实现奠定了基础。随着算法规制成为数字法治核心议题,数据正义向算法治理领域纵深发展,围绕算法治理的权益分配、是否可信、能否审核、怎样追溯、何以救济等问题形成了一种关于算法公平的学说。

1. 算法公平的内涵

算法公平包括结果公平和过程公平,前者聚焦算法产生法律后果之后,后者侧重算法产生法律后果之前。两者均强调以算法技术提升犯罪治理能力的同时,不能对国民权利造成超出法律限度的减损。

结果公平是指算法引发法律后果的公平性,源自数据正义的反对不公平对待原则,从对算法歧视、偏误等不公平对待的回应入手,聚焦算法决策结果及其影响是否准确且正当,防范算法引发、加剧对相关群体的歧视,强调算法偏误的可救济性和可问责性。具言之,犯罪治理算法的结果公平以算法结论的精准性与稳健性为前提,通过有效的损害救济程序与清晰的责任制度消除明显不公,推动各方主体的权利义务关系向着更公平方向发展。这是评价算法公平的基本要求,若基本要求未达到,相关算法将面临暂停或终止的命运,相关责任人亦可能面临被追责的风险。以地点评分为例,“在经历了当地多年抵制以及关于种族歧视和公民自由问题的全国性批评后,洛杉矶市警察局于2020年春季终止了运行长达十年的PredPol算法预测系统实验,该实验旨在预测侵犯犯罪的高发地点”^②。与之类似,英国卡迪夫大学“数据正义实验室”2022年发布报告指出,美国、澳大利亚、加拿大等国的61个基于算法决策的政府项目被终止或暂停,这些项目多来自福利欺诈调查、警务与法律等领域,取消原因有算法的有效性存疑、社会批评及对隐私、公平、偏见、歧视的忧虑等。^③

过程公平指向算法运行过程的公平性,源于数据正义的可见性和技术参与原则。在犯罪治理中,个体意义上的平等与自主权在一定程度上需通过制度层面、整体意义的法律授权来实现。因此,过程公平既强调在自动化决策中保障个体的程序性权利,“确保对特定个体裁决和规则制定的公平、透明和问责”^④;也涵盖算法影响更广泛主体权利的

^② See Linnet Taylor, *What is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally*, Big Data & Society, Vol.4:1, p.1, 9 (2017).

^③ Jeremy Davis et al., *Five Ethical Challenges Facing Data-driven Policing*, AI and Ethics, Vol.2:185, p.185 (2022).

^④ See Joanna Redden et al., *Automating Public Services: Learning from Cancelled Systems*, at <https://carnegieieuk.org/publication/automating-public-services-learning-from-cancelled-systems/> (Last visited on August 1, 2025).

^⑤ Danielle K. Citron, *Technological Due Process*, Washington University Law Review, Vol.85:1249, p.1251 (2008).

过程,注重算法运行的正当合法、公开透明、科学高效。犯罪治理算法的过程公平包括如下正当性要求:其一,算法治理的启动应遵循职权法定,法律授权乃算法治理的合法性依据。其二,算法运行应具备合比例性。尽管算法控制是犯罪治理之必需,但要尽量避免算法过度干预社会生活和造成不必要的损害。其三,算法治理应恪守正当程序原则,以算法公开等程序保障算法透明度和可解释性、增进算法信任与约束技术滥用。这些正当性要求确立了比结果公平更高的法治追求。犯罪治理算法的过程公平与否,关键要看在追求安全目标的过程中,算法权力扩张是否与个人权利保障相匹配,算法权力有无运行于法治轨道,算法风险能否受法律规制,算法使用者与受算法影响者之间的“权力—权利”关系是否处于均衡状态。

总之,结果公平和过程公平均聚焦于算法权力的规范化和法治化,回应并预防算法运行中威胁隐私、自由、人格尊严的潜在风险,追求算法治理的系统性正义。二者共同构成算法公平的内涵,亦成为评判犯罪之算法治理是否正当、合理的价值坐标。

2. 算法公平的实现

算法权力的规范有赖于算法规制体系的建构,围绕结果公平和过程公平这对范畴,形成了回应性规制和预防性规制两种模式。回应性规制通过算法救济及时回应受算法影响者乃至公众的申诉和控告,恢复因算法偏误或歧视而造成的负面影响;通过算法问责追究算法使用者不当使用的回顾性法律责任,确保算法治理的结果公平。预防性规制通过算法影响评估和算法公开等机制,要求算法使用者承担更全面的法定义务,确立规范算法权力的前瞻性法律责任,保障算法治理的过程公平。对回顾性责任的讨论是通过观察损害结果或算法偏见来展开的;对前瞻性责任的讨论意味着算法使用者有责任考虑算法的潜在影响,并采取措施防止损害发生。^{②⑤}然而,相对于商业算法,犯罪治理算法的法律规制尚处于探索阶段。公权力机关和看门人市场主体的犯罪治理职责具有法定性和强制性,致使犯罪治理算法应用及算法权力扩张偏重“监视—筛选”的行动逻辑,而对“回应—预防”的规制逻辑重视不足。对此,应完善算法治理在“回应—预防”规制逻辑上的法律制度,探索“依据法律的规制”。《网络安全法》《个人信息保护法》《反电信网络诈骗法》等虽为算法规制提供了一定依据,但尚缺乏针对公共算法尤其是犯罪治理算法的系统性规定。从规制逻辑看,国家机关或互联网平台的算法控制应以法定职责或法定义务为依据;即便某些算法先行先试是出于治理的迫切需要,也应适时推动法律制度的创制和续制,做到算法治理于法有据、循法而治。同时,“依据法律的规制”须以“针对算法权力的规制”为正当性内核,以回应性规制和预防性规制规范算法权力,保障受算法影响者乃至全体国民之合法权益。

(二) 算法偏误对结果公平的侵蚀

犯罪之算法治理取得了显著效果,但对标算法公平原则发现,某些算法偏误难以获

^{②⑤} See Marie C. Decker, Laila Wegner & Carmen Leicht-Scholten, *Procedural Fairness in Algorithmic Decision-making: The Role of Public Engagement, Ethics and Information Technology*, Vol.27:1, p.9 (2025).

得及时纠偏,导致侵蚀结果公平。算法公平并非追求乌托邦隐喻中的绝对公平,而是致力于减少对受算法影响者的明显不公平对待。实践中,算法时常因不完备的训练数据、不科学的模型乃至人类固有偏见等原因出现偏误。例如,美国用于再犯风险评估和预测的“罪犯矫正替代性制裁分析管理系统”(COMPAS)由于训练数据隐含种族偏见,导致裁判中出现不少显失公正的问题。^{②6}芝加哥警局用于预测枪击暴力犯罪的算法系统——“战略目标名单”(Strategic Subject List),对近四十万名曾被捕的个人进行涉枪暴力风险预测评分,但不少证据表明存在严重的算法歧视和偏误。^{②7}“断卡”行动中算法风控频繁“误伤”用户,不少人被银行误判为高风险用户,被采取限额、止付、冻结等措施,给用户接受金融服务带来不便。因此,通过救济和问责程序纠正算法偏误实属结果公平的应有之义,但囿于法律依据不完善,算法救济和问责的实现存在诸多困难:

在救济方面,算法救济制度在广泛场景中的普适性不足,存在相当程度的法律空白,导致受算法影响者提出异议、申请纠正的权利难以获得保障。以电诈犯罪的算法治理为例,《反电信网络诈骗法》第32条第3款规定,“被处置对象可以向作出决定或采取措施的部门、单位提出申诉。作出决定的部门、单位应当建立完善申诉渠道,及时受理申诉并核查,核查通过的,应当即时解除有关措施”。遗憾的是,仅凭《反电信网络诈骗法》之原则性规定尚不足以有效落实算法救济的各项要求,算法使用者的告知、核查和纠正等程序亟待完善细化。同时,反电诈的算法救济机制也无法直接适用于其他类型的网络犯罪治理,其他领域的算法治理更缺乏明确详尽的救济程序。

在问责方面,算法责任追溯面临问责机制缺失困境及因果关系模糊、责任边界不清等责任认定难题。算法治理不具有绝对的技术“中立性”,精密齿轮般运行的算法之基底是算法权力的流淌,对造成严重后果的算法仅靠救济不足以保障结果公平,还应有对算法的终止、对算法使用者的问责等事后追责机制。对存在严重问题的算法进行干预相对容易实现,而对算法使用者的问责尚无完备的法律依据。尽管《个人信息保护法》第7章和《反电信网络诈骗法》第6章中的责任条款规定,存在违反个人信息处理规定或给他人造成损害等情况时,相关主体应依法承担法律责任,但遗憾的是在犯罪之算法治理领域尚未形成算法问责的一般性规则。

(三) 算法失范对过程公平的偏离

随着预防性治理的勃兴,公权力机关对违法犯罪的算法风控呈扩张趋势。人脸识别、步态识别、内容审核、敏感数据推断等对个体渗透度强、个人信息使用度高的技术措施均以算法为底层逻辑,犯罪治理中遍布算法风控身影。然而,犯罪之算法治理的专门性法律依据相对滞后,或缺乏明确的法律规定,或仅被法律、政策概括性提及。例如,反电诈的算法风控隐含于《反电信网络诈骗法》之反制技术条款中;维稳风控在各省《平

^{②6} See Sara M. Smyth, *Can We Trust Artificial Intelligence in Criminal Law Enforcement?*, Canadian Journal of Law and Technology, Vol.17:99, p.107-108 (2019).

^{②7} See Andrea L. DaViera et al., *Risk, Race, and Predictive Policing: A Critical Race Theory Analysis of the Strategic Subject List*, American Journal of Community Psychology, Vol.73:91, p.91 (2024).

安建设条例》等法规中仅原则性地表述为“推动数字技术与矛盾纠纷化解相结合”“推动人工智能与平安建设的深度融合”；大数据反腐在中纪委公报中表述为“以大数据信息化赋能正风反腐”^{②⑧}。在一般法层面，《个人信息保护法》为国家机关运用算法处理个人信息设定了合法正当、最小必要、合目的性、公开透明等原则性要求，虽然“采用了公民法一体调整的立法体例……但相关内容过于原则和简单，诸如国家机关处理个人信息的正当性基础、信息处理的具体规则、信息主体享有的权利、国家机关违法处理个人信息应当承担的责任等一系列关键问题的规定均存在不明确的地方”^{②⑨}。一般性规定过于原则和专门性规定相对薄弱，导致偏离过程公平的算法失范时有发生，具体表现在：

第一，算法治理与职权法定的正当性要求存在差距。“就预测对象而言，预测措施所涉及的相对人并非仅仅是潜在犯罪人，被害人、普通公众都有可能被纳入预测的视野。”^{②⑩}涉个人信息的算法治理应以算法使用者的法定职责或义务为限。而公权力机关开展算法治理存在法定职责不明晰的问题，算法应用的范围、程序、权限等缺乏规定。对刑事立案的案件，算法引导侦查具有实定法依据，但立案前的算法风控尚无规定。如大数据扫黄应限于组织、强迫卖淫等刑事犯罪侦查且在立案后开展，在立案前的线索经营阶段允许有限制地以算法固定证据，进而促成立案。然而，仅对卖淫嫖娼的治安违法开展算法风控，须以《治安管理处罚法》《公安机关办理行政案件程序规定》为限，但相关法律并无算法使用的明确授权。这无疑与“法无明文规定不可为”的精神相抵牾，导致算法风控处于失序状态。再如，大数据反腐包括无涉个人信息的算法风控和涉个人信息的算法风控。前者以公权力运行轨迹为对象，可用于纪检监察各领域；涉个人信息的算法风控应主要限定于立案审查调查的职务犯罪。可惜，对于违纪查处、巡察预画像、日常监督的算法风控能否以及如何使用被监督对象的个人信息尚缺乏明确规定。

第二，算法治理有时偏离比例原则，对个体生活造成过度干预。大数据扫黄具有保护妇女权益、铲除涉黄黑恶势力的作用；但治安防控不能以效率为唯一目标，依托“旅馆业治安管理信息系统”的算法风控涉嫌对一般公众的涉黄风险开展大数据评分。此举偏离了禁止权力过度扩张的比例原则，非保护公共利益之必要，易陷入过度管控的桎梏。涉个人信息的算法治理还可能引发寒蝉效应，“人们进行自我审查并回避某些活动，未必源于对被起诉的恐惧，而是出于对潜在未来损害的担忧，这些损害可能来自隐私泄露、公开披露、欺诈或身份盗用风险，或者被公权力机关贴上犯罪、越轨者和行为不端者的标签”^{②⑪}。维稳风控对相关群体的预警及排查，有助于推动矛盾纠纷的实质化解和重大风险的源头预防，但亦潜藏着对相关人群标签化、超权限处理相关人员个人信息的风险。

^{②⑧} 参见《中国共产党第二十届中央纪律检查委员会第四次全体会议公报》，2025年1月8日中国共产党第二十届中央纪律检查委员会第四次全体会议通过。

^{②⑨} 前注^{②⑩}，马怀德文，第18页。

^{②⑩} 蒋勇、张晓华：《算法时代预测性警务的兴起及其风险规制》，载《公安学研究》2023年第2期，第12页。

^{②⑪} Jonathon W. Penney, *Chilling Effects: Online Surveillance and Wikipedia Use*, Berkeley Technology Law Journal, Vol.31:117, p.164 (2016).

第三,算法治理时常因黑箱问题有违正当程序。“警务算法实际上形成了自审、自用、自评估的封闭循环体系,现有算法监管体系很难加以干预……算法权力隐藏于组织行为之中更加难以被识别。”^{③②} 为保护公共安全和保守商业秘密,犯罪治理算法在数据收集、模型构建、结果输出等环节体现出高度的自主性和封闭性,形成了一种有意为之的不透明。公众既无法了解算法如何得出具体结果,也无从判断决策过程是否正当合理。“算法自动化决策的不透明性遮蔽了行政公开原则,算法设计运行的技术垄断架空了公众参与原则,算法决策黑箱无法为相对人提供理由说明,算法自动化决策与行政正当程序产生了严重冲突。”^{③③}

综上,在“以算法为中心的智能社会法律秩序”^{③④}中,必须防范算法工具属性对以人为本价值的背离,将算法公平的正当性要求投射到算法规制体系之中。“人类历史表明,任何一次巨大的技术革命,都会带来总体性的制度变迁和范式转型。”^{③⑤} 由于犯罪之算法治理兴起不过十余年,相关规制体系建构尚处于起步阶段。对此,可行的思路是以算法公平的实现为导向,沿着回应性规制和预防性规制路径约束算法权力,通过算法救济和问责等回应性机制保障结果公平,依托算法影响评估和算法公开等预防性机制追求过程公平,以此建构兼顾结果公平和过程公平的算法规制体系(详见图1)。

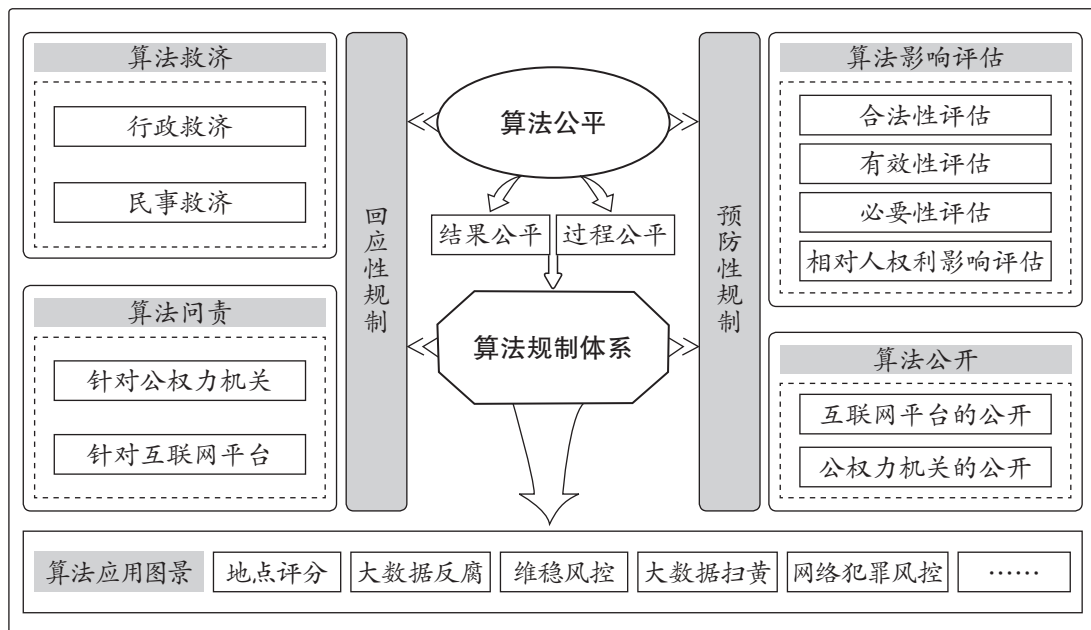


图1 犯罪治理算法的规制体系

③② 蒋勇:《算法时代警察权的隐性扩张及其应对》,载《苏州大学学报(法学版)》2024年第1期,第63-64页。

③③ 张凌寒:《算法自动化决策与行政正当程序制度的冲突与调和》,载《东方法学》2020年第6期,第7页。

③④ 张文显:《构建智能社会的法律秩序》,载《东方法学》2020年第5期,第4页。

③⑤ 马长山:《数字法学的理论表达》,载《中国法学》2022年第3期,第120页。

三、犯罪治理算法的回应性规制

鉴于犯罪治理算法的偏误有时难以避免,回应性规制必须关注算法引致的负面影响,依托算法救济和算法问责机制保障结果公平。其中,算法救济指向受算法影响者的权利恢复;算法问责则是对算法本身的干预及对算法使用者的追责和监督。两者分别从算法治理的对象端、主体端作出复权和限权导向的回应。

(一) 回应权利受损的算法救济

算法救济是对算法自动化预测结论的人工干预和修正。犯罪治理算法的应用关涉受影响者是否进入执法或侦查的视野,关系到受影响者的人身、财产等权利是否受限受损,在权利保障方面的高风险性使其与商业算法存在本质区别。犯罪之算法治理包括前后衔接的预测与决策阶段,并在这两个阶段之后相应地存在具有纠正算法偏误、修复受影响者权益作用的两道救济防线。其一,在预测阶段之后,自动化预测的结论须臾离不开决策阶段的人工评估与修正。试图让算法纤悉无遗且完全准确地预测违法犯罪风险并不现实,未经决策阶段的人工干预而径直将预测结论作为犯罪控制依据,可能引发算法偏误,故而对预测阶段自动化分析结论进行基于客观事实的人工评估实属保障结果公平的第一道防线。其二,在决策阶段之后,即便经历人工评估,囿于基础数据不客观、风控规则不合理、系统设计不科学等原因,可能依然存在算法决策不准确问题,对受影响者申诉的核查及救济成为保障结果公平的第二道防线。面对难以完全避免的算法误判,须针对不同的行为主体,从行政救济与民事救济两个方面予以完善。

从行政救济看,应完善犯罪之算法治理的“告知—申诉”与“核查—反馈”机制。“对于算法公权力的应用,应提前告知相对人和利益相关人;对于有异议的情况,应提供制度化的投诉和救济渠道。”^{③⑥}《电信网络诈骗及其关联违法犯罪联合惩戒办法》(以下简称《办法》)虽非直接针对算法风控,但通过设置事先告知、多元申诉、限期核查、书面反馈和错误纠正等程序,细化了《反电信网络诈骗法》第31条第2款的申诉救济条款,为犯罪治理算法的行政救济提供了借鉴范本。算法偏误虽不会直接引发承担刑事或行政责任的终局性法律后果,但却能在一定时间内对治理对象的正常生活及相关权益造成实质性影响。“告知—申诉”与“核查—反馈”机制的细化完善,有助于及时纠正对受影响者合法权益的不当干预,保障正常生产经营和群众生活便利。在未来的人工智能立法及算法规制的制度设计中,可参考《办法》与《反电信网络诈骗法》第32条第3款之规定,在更广泛的犯罪治理领域确立以“告知—申诉”与“核查—反馈”为核心的算法行政救济制度。在“告知—申诉”机制中,应明确依据算法作出决定或采取措施的执法部门之告知义务和告知时限,在以书面形式向受影响者披露决策事由、依据、影响范围等事项的基础上,还应告知受影响者所享有的申诉权利和申诉方式,让受影响者可根据

^{③⑥} 周辉:《算法权力及其规制》,载《法制与社会发展》2019年第6期,第125页。

告知内容,通过线下线上等多元方式提出申诉。在“核查—反馈”机制中,应明确作出决定或采取措施的部门开展实质审查的标准、时限及反馈方式。对维持算法决策结论的,应向受影响者书面说明理由;对决策结论存在偏误的,应启动撤销错误标记、解除关联限制、恢复受损权益等补救措施,并对由此造成的损害承担赔偿责任。

从民事救济看,互联网平台、金融机构等市场主体的算法风控对用户造成不当限制或损害的,受影响者有权向平台申诉,或提起民事诉讼要求市场主体承担侵权或违约责任。市场主体的内部申诉机制既是可供用户选择的便捷救济途径,也是依据平台规则或企业制度开展的内部核查。《反电信网络诈骗法》第32条第3款对市场主体建立、完善申诉渠道和核查机制提出要求。当下,用户的民事救济多通过向作为算法使用者的市场主体申诉实现,但内部申诉有时面临“制度设计不科学、程序中立性不足、解纷人员专业性公信力不强、与执法司法衔接不畅等问题”^{③7},仍需具有外部性的民事诉讼予以补充。用户可基于与市场主体的合同关系,要求其承担继续履行、采取补救措施或者赔偿损失等违约责任;也可基于侵权事实要求市场主体承担停止侵害、排除妨碍、赔偿损失、消除影响等侵权责任。实践中,用户因电信运营商、银行采取算法风控涉嫌侵权而提起诉讼的案件频发,法院的裁判既有支持用户损害赔偿请求的判例,也有驳回诉讼请求的情形。争议的症结在于,与纯粹的民事法律关系不同,当市场主体作为看门人承担犯罪治理职能时,用户对算法控制负有一定容忍义务,但容忍义务的边界尚存在模糊地带。必须承认,容忍义务不等于豁免责任。在市场主体运用犯罪治理算法履行安全防卫职责时,需根据市场主体是否及时解除风控措施、决策合理性、处置规范性等事项,具体判断是否存在过错、因果关系和免责事由。

总之,人工智能立法应重视包括犯罪治理算法在内的算法救济一般性条款的制度供给,从行政救济和民事救济上明确算法侵权的申诉程序和赔偿责任认定规则,保障受影响者的合法权益和算法结果公平。

(二) 回应权力致损的算法问责

不同于面向权利恢复的救济,算法问责聚焦对算法致损后果的责任认定和追究,“主要解决算法风险产生后责任的可追溯性、一般配置及具体分担问题”^{③8},围绕特定场景明确算法使用者等主体的责任,以实现对其权力的控制。关于犯罪治理算法的问责,应依据算法使用主体的不同,区分为公权力机关与市场主体两类予以探讨。

一方面,当公权力机关的犯罪治理算法存在严重偏误等问题时,其责任追究方式既包括针对算法的及时修正或责令终止等矫正性措施,也包括针对相关责任人员的党纪政务处分等惩戒性措施。因犯罪之算法治理呈现强烈的“实验主义治理”^{③9}色彩,具有探索性和实验性,故算法问责在更多场景下聚焦施行阻力小且反馈敏捷的矫正性措施。在循证检验中发现算法存在不可接受的风险时,应以矫正性措施及时予以干预与修正。

^{③7} 黄文艺、孙喆明:《论互联网平台治理的元规制进路》,载《法学评论》2024年第4期,第120页。

^{③8} 苏宇:《算法规制的谱系》,载《中国法学》2020年第3期,第174页。

^{③9} 林涓民:《论人工智能立法的基本路径》,载《中国法学》2024年第5期,第100页。

在域外,针对算法的矫正性措施已被用于规制人脸识别及福利欺诈监测等算法。2018年,某社会组织将荷兰政府用于福利欺诈风险监测的“系统风险指示”(System Risk Indication/SyRI)诉至法院,指责人工智能系统存在算法不透明和过度监视问题。海牙法院认为,该系统没有足够的措施防范隐私风险、工具运行缺乏透明度、可能因社会经济或移民身份而产生歧视,并以违反人权规范为由裁定于2020年终止使用。^{④①}2019年,针对警员的随身摄像头在街面对公众无差别扫描的疑虑,美国加利福尼亚州旧金山市通过《停止秘密监控条例》,成为美国首个禁止政府部门使用人脸识别技术的城市。^{④②}随后加利福尼亚州通过《AB 1215 法案》,禁止州和地方执法部门的警用随身摄像头使用人脸识别。^{④③}与之相对,惩戒性措施是根据算法致损的事实、情节、主观过错等因素,对算法使用者实施行政追责和纪律处分。囿于算法的复杂性、封闭性和探索性,当犯罪治理算法引发负面影响或造成损失时,如果算法使用者在合理的时间内采取矫正性措施而停用或修正算法,后果被及时控制且尚未扩散,并经算法解释和算法审计表明算法致损是不能预见或无法避免的,则属于在现实条件下履行了审慎注意与风险规避义务。否则,应根据犯罪治理算法在数据质量、系统设计或判别规则等方面的缺陷,以及算法使用者的主观过错,判定相关责任人员就算法致损后果以及后果的扩大等负面影响承担行政或党纪责任。在未来的人工智能立法中,应明确对算法致损的矫正性措施和惩戒性措施两种问责方式。

另一方面,当互联网平台等市场主体不当使用算法导致受影响者受损时,其责任追究应基于看门人义务履行与受影响用户权益保障的均衡立场,以前置化的行政责任为规制路径。在看门人规制中,犯罪治理算法的致损具有鲜明的累积性特征,其损害形态包括数据不当处理引发的用户信息权益侵害、错误风险标签导致的区别性对待与名誉权受损、资金管控不当造成的财产权受损等。如果仅孤立地评价,其对特定用户的损害相对有限,但经网络规模化效应的累积,则可能严重影响网络空间的预防性秩序建构,妨害监管部门和用户对人工智能系统及其背后看门人的信赖等公共利益。行政法律责任的设置可实现对此类算法侵害的主动发现、调查与处置,弥补民事救济的被动性与滞后性。有鉴于此,应充分发挥已有制度依据的作用,将互联网平台等主体在犯罪治理中违规处理个人信息与数据的行为纳入行政规制框架。《网络安全法》《个人信息保护法》等法律已建构起覆盖自动化决策的问责框架,对违规处理个人信息或数据的行为,明确了包括责令改正、警告、没收违法所得、罚款、责令暂停或终止相关业务、停业整顿、关闭网站、吊销相关业务许可证、吊销营业执照及处罚责任人在内的一揽子行政责任。同时,有必

^{④①} 参见《联合国专家:荷兰法院里程碑式的裁决将制止政府对穷人进行数字监视》,载联合国官网, <https://news.un.org/zh/story/2020/02/1050372>, 2026年2月1日访问。

^{④②} See Mike Maharrey, *First in the Nation: San Francisco Passes Ordinance to Ban Facial Recognition, Tenth Amendment Center*, at <https://blog.tenthamendmentcenter.com/2019/05/first-in-the-nation-san-francisco-passes-ordinance-to-ban-facial-recognition/> (Last visited on January 24, 2026).

^{④③} See Rachel Metz, *California Lawmakers Ban Facial-recognition Software from Police Body Cams*, at <https://edition.cnn.com/2019/09/12/tech/california-body-cam-facial-recognition-ban/index.html> (Last visited on January 26, 2026).

要在立法层面明晰犯罪治理算法致损的行政处罚规范及豁免条件。必须承认,市场主体在履行犯罪控制义务过程中已承受了成本提升与用户流失的双重压力,若让其承担过度严苛的行政责任,可能对看门人正常经营造成负面影响。2025年12月1日施行的《浙江省实施〈中华人民共和国反电信网络诈骗法〉办法》第39条第2款为看门人和其他履行反电诈职责的单位及其工作人员确立了“尽职豁免”规则,即“因依法履行职责、义务被投诉,或者发生电信网络诈骗案件但有证据证明其已经尽职尽责的,不予追究责任”。对于依法履行算法救济、算法影响评估、个人信息保护等合规义务的互联网平台等市场主体,可依循上述条款免受算法致损的行政追责。

上述两点分别探讨了公权力机关与互联网平台开展算法治理时各自所应承担的责任,但基于技术支持与协助执法义务的履行,互联网平台时常与公权力机关开展协同治理,对协同治理中的算法致损问责,则应根据各自职责范围、义务边界和原因力比例划分责任,按照各主体不当使用算法行为的事实、性质、情节及严重程度分别认定。

四、犯罪治理算法的预防性规制

回应性规制属于对犯罪治理算法的事后规制,是约束算法权力的基本要求;预防性规制是对算法的事前规制,以算法影响评估和算法公开等方式保障数字时代的国民权益,从整体上控制算法风险、提升算法透明度并推动算法运行于法治轨道。预防性规制对过程公平的追求构成了对算法规制的更高要求。犯罪治理算法因其特有的高度敏感性,决定了算法影响评估更多体现为一种内部规制,而以透明度为核心的算法公开系典型的外部规制。

(一) 依托算法影响评估的内部规制

算法影响评估以风险防范为目标,于事前审查算法应用是否符合职权法定、比例原则、正当程序原则,以缓解事后问责的迟滞并防范风险扩散。相对算法救济和问责的被动性,算法影响评估具有鲜明的主动控制、过程控制和内部控制色彩。《个人信息保护法》第55条规定利用个人信息进行自动化决策时,“个人信息处理者应当事前进行个人信息保护影响评估”。参考《个人信息保护法》第56条对个人信息保护影响评估内容的规定,犯罪治理算法的影响评估至少包括以下四个方面:

第一,合法性评估。合法性评估要求对犯罪治理算法的启动理由、主体、对象、范围、程序等事项是否具有完善的法律授权进行审查。算法虽然显著增强了治理能力,但“在公共服务中部署人工智能的风险极高,直接触及公众信任与治理合法性等公共权威赖以存在的根本性基础”^④。“为避免个人尊严的保护被完全淹没在追求技术福利的目标之下,对公共决策适用算法设置界限无疑是必需的。”^④根据“法无授权不可为”的法治精神,

^④ Johann Laux, Sandra Wachter & Brent Mittelstadt, *Trustworthy Artificial Intelligence and the European Union AI Act: On the Conflation of Trustworthiness and Acceptability of Risk, Regulation & Governance*, Vol.18:3, p.5 (2024).

^④ 赵宏:《公共决策适用算法技术的规范分析与实体边界》,载《比较法研究》2023年第2期,第7页。

合法性评价应以算法应用是否具备法律依据为准。考虑到算法创新多具试验性特征,在探索阶段以“缓和的职权法定原则”为评估标准颇为现实。有学者指出,“在改革语境下,职权法定原则仍需被遵守,但不得有所缓和……一旦确认改革制度的有效性,要固定或推广改革制度,就需要立即补足合法性……虽然在改革时,职权法定原则要有所缓和,但这种缓和本质上是法律的一种暂时的、局部的空缺或悬置,主要是缓和时限,并非缓和规范强度”^{④5}。缓和的职权法定说契合了算法创新与法治保障相均衡的需要,构成了合法性评估的原理。网络犯罪的算法风控就经历了从实践先行到制度跟进的缓和职权法定的演进历程。互联网平台最初基于落实主体责任的政策要求,以算法创新回应犯罪挑战;在算法风控成熟后,《反电信网络诈骗法》等制度逐步完善了看门人的犯罪控制义务及算法治理依据。可见,算法规制的法律建构并不是一蹴而就的,盲目追求刚性的职权法定更不切实际,合法性依据的完善有一个从治理到制度、从原则到具体、从框架到细则的建构过程。

第二,有效性评估。有效性评估是指在算法设计、部署阶段以及运行中持续进行的治理效果评判,其发生于算法投入使用前和运行中,包括检验算法训练数据的科学性,分析原理及参数设计的合理性,考察算法决策的可解释性和可检验性,探究风控对象画像的客观性和可靠性,评估算法效果的准确性等。如果某种算法在运行中被评估为效果不佳、无法实现目标,应及时终止该算法,并分析原因、留存记录;如果某种算法在某些环节或场景具备有效性,而在其他方面又存在缺陷,则应修正算法并再次验证修正后的效果。在域外,“由于预测性警务很大程度上未经实践检验,各执法部门必须自行评估初步成效的可靠性。旧金山警察局曾对引入预测性警务的可靠性进行评估,但最终因对实际效果存疑而放弃采用 PredPol 技术”^{④6}。是故,需定期评估算法模型的设计参数、评分规则、预警效果,系统性识别并修正潜在偏差,确保预警结果的准确性与公正性,以实现算法公平的预防性保障。

第三,必要性评估。必要性评估是在部署算法之前评判其对个体权利可能引发何种威胁,并厘定该算法是否是达成目标之必需且为威胁个体权利最低手段的过程。人工智能系统的风险分级为必要性评估提供了分析框架。欧盟《人工智能法案》对可能侵害基本权利的人工智能系统采取风险分级监管,分为禁止性、高风险、有限风险、最小风险四类。我国《生成式人工智能服务管理暂行办法》《互联网信息服务算法推荐管理规定》等规章虽提出对生成式人工智能服务、算法推荐技术实行分类分级管理,但尚未形成系统化的算法分级制度。应根据应用场景的敏感性、决策后果的重要性、技术干预的强度等因素,将犯罪治理算法纳入风险分级分类管理。在此基础上,针对不同风险等级,匹配相应强度的必要性评估。评估的重点包括算法保护法益的重要性、有无风险更低的替代方案及运行成本是否合理等因素。欧盟《人工智能法案》第5条规定实时远程生物识

^{④5} 白云锋:《行政执法权下放法律依据之建构》,载《现代法学》2025年第2期,第161-162页。

^{④6} Andrew G. Ferguson, *Policing Predictive Policing*, Washington University Law Review, Vol.94:1109, p.1159 (2017).

别系统等属于被禁止的人工智能应用,但出于维护公共安全和打击严重刑事犯罪之必要,也允许在严格限定的条件下使用,以实现社会安全防卫与个体权利保障的平衡。在维稳算法的必要性评估中,因算法风控对象不仅有犯罪嫌疑人,还包括一般违法以及尚未违法但有一定风险的行为人(如扬言行凶报复社会、多次滋事、缠访闹访的行为人),维稳算法的应用不能过度扩张,应主要聚焦防范“民转刑”“刑转命”等极端恶性案件,发挥“挽狂澜于既倒”的功效。

第四,相对人权利影响评估。此评估专注于算法对所影响的个体或群体基本权利产生的现实损害或潜在威胁,且被欧盟多部法律及政策文件确认。欧盟《通用数据保护条例》第35条提出处理与刑事定罪和犯罪相关的个人信息应进行数据保护影响评估的要求。欧盟《人工智能法案》第27条规定,在部署高风险人工智能系统前,部署者须对之进行全面评估,评估内容涵盖具体程序、使用期限与频率、受影响自然人和群体类别、具体危害风险、人工监督措施实施情况及风险应对等内容。鉴于犯罪治理算法不可避免地隐私权、人身权等实体性权利与知情、申诉等程序性权利产生重大影响,相对人权利影响评估由此成为一种增进社会信任不可或缺的强制性要求。对于大数据反腐的权利影响评估,应依据新修订《监察法》“尊重和保障人权”之基本原则,不仅应关注受算法影响的被监督对象的隐私、个人信息等实体性权利,还应考察算法运行透明度与权利救济方式等程序性权利,并将这些要求体现于监察法规及相关党内法规的完善之中。

上述四项机制构成了算法影响评估的基本内容,为犯罪治理算法合规运行提供了法治标准。同时,还需要明确评估主体和评估频率,以进一步建构针对犯罪治理算法的影响评估体系。从评估主体看,犯罪治理算法具有较强的封闭性和保密性,算法部署者和使用者无疑是有责任、有能力开展评估的适格主体,故而算法影响评估主要体现为内部评估;但在涉公共安全、隐私保护等高风险场景时,外部评估的独立性亦有重要价值。2019年发布的欧盟《算法责任与透明治理框架》指出,应将公共机构的自我评估、公众参与及外部研究者访问整合为一个过程,保证算法引入前的影响评估,为征求和处理公众意见提供机会。^{④7}从评估频率看,参考2017年欧盟委员会《关于数据保护影响评估指南(DPIA)和为GDPR目的确定处理活动是否“可能导致高风险”的指南》的相关要求,“数据保护影响评估须在处理之前进行,但开展数据保护影响评估是一个持续的过程,而非一次性活动”^{④8}。可见,犯罪治理算法的影响评估亦非一次性的合规任务,而是一个与算法迭代相伴的持续性规制过程,贯穿算法运行全生命周期,包括初始评估、定期评估、事件驱动评估等节点。算法在投入运行前需完成首次影响评估,此后的定期评估频率与算法风险等级挂钩。当算法本身及外部环境发生重大变化时,则触发再次评估。

^{④7} See Ansgar Koene et al., *A Governance Framework for Algorithmic Accountability and Transparency*, at [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU\(2019\)624262_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf) (Last visited on February 13, 2026).

^{④8} *Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is “Likely to Result in a High Risk” for the Purposes of the GDPR*, at <https://qubitprivacy.com/wp-content/uploads/2021/06/qubit-briefing-6-dpia-v3.pdf> (Last visited on October 17, 2025).

（二）基于算法公开的外部规制

犯罪治理算法离不开对治理对象个人信息的使用,算法使用者与治理对象之间往往呈单向度透明态势,“其中一方是深藏不露的观察者和控制者,另一方则是浑然不觉的被观察、被控制的对象……这无疑会导致严重的数字不公平,亟待予以制度性解决”^④。为此,“建立透明、可参与的流程以审视算法渗透型社会现实”^⑤,改善算法透明度成为算法规制的关键。这一正当要求可概括为算法公开及可解释性。基于《个人信息保护法》及相关政策文件提供的初步制度基础,^⑥算法公开及可解释性强调将算法及其使用情况以易于理解的方式对社会告知或说明,公开算法名称及算法使用者信息、算法的功能与目标、数据的来源和类型、自动化决策的原理等事项,以保障算法透明度和国民知情参与权利。

当前,犯罪治理算法公开面临三重限制:一是即便公开算法全部细节,公众仍无法理解;二是如何防范不法分子利用算法公开规避打击;三是算法公开不能违背保守国家秘密、工作秘密和商业秘密之法律规定。三重限制意味着犯罪治理算法的公开是保障可理解性的有限公开,且存在一定范围的公开豁免空间。在坚持算法透明的前提下,应当允许在特定情形下豁免公开某些算法的全部或部分信息。关于算法公开的思路,有学者将算法运行分为现象层和隐藏层,提出算法“分层规制”策略,主张针对现象层进行有限公开。^⑦因此,犯罪治理算法的公开亦为基于算法解释的有限公开,解释的重点并非算法源代码,而在于解释算法原理和决策过程,并公开责任主体。《互联网信息服务算法推荐管理规定》明确了互联网企业部署和使用算法的公开说明义务、备案义务,为互联网企业的犯罪治理算法公开提供了制度依据。相对互联网企业的算法公开,公权力机关的算法更应以适宜方式备案或登记,向社会告知算法的目标、数据来源、决策原理和运行机制,说明算法应用的必要性,披露算法运行效果等。以往法律之所以未规定犯罪治理算法公开,不是不能公开,而是法律发展有一定滞后性,算法公开制度的确立有一个从互联网企业向公权力机关延展的过程。

1. 互联网平台的算法公开

在算法备案及透明性要求下,互联网平台的犯罪治理算法公开取得了突破性进展。平台通过算法备案、平台规则公开等方式在一定程度上落实了犯罪治理算法的公开及解释义务。

第一,互联网平台在“互联网信息服务算法备案系统”中公开。算法备案系统极大提升了算法运行的透明度和可解释性,保障了对算法应用的外部监督。犯罪治理算法在

^④ 前注^③,马长山文,第136页。

^⑤ Claudia Wagner et al., *supra* note 1, 202.

^⑥ 《个人信息保护法》第7条规定处理个人信息的公开、透明原则;第17条规定个人信息处理者应向个人告知或公开处理规则;第48条规定个人信息处理者的解释说明义务。2021年9月国家网信办等九部委印发的《关于加强互联网信息服务算法综合治理的指导意见》指出,应推动算法公开透明,督促企业及时、合理、有效地公开算法基本原理、优化目标、决策标准等信息,做好算法结果解释,畅通投诉通道,消除社会疑虑。

^⑦ 参见蔡星月:《算法规制:从规范规制到分层规制》,载《现代法学》2024年第4期,第187-189页。

该系统中主要被归为“检索过滤类”或“内容过滤类”,如阿里、百度、腾讯、快手等公司的“内容安全算法”,京东公司的“内容过滤算法”等,备案和公开的内容包括算法名称、基本原理、运行机制、应用场景、目的意图等事项。

第二,互联网平台在隐私政策及其他平台规则中公开。《支付宝隐私权政策》提出,“为保障服务的稳定性与安全性,保护您的账户与资金安全,履行反洗钱等法定义务,我们会将您的信息用于身份验证、安全防范、诈骗/盗用/可疑交易监测、预防或禁止非法活动、降低风险及防止风险传导、存档和备份用途”^{⑤③}。《淘宝网算法原理及信息处理情况说明》对“淘宝内容安全算法”的算法原理、运行机制、应用场景、算法目的等事项作出了详细说明。

第三,互联网平台开通算法公开专区。抖音建设安全与信任中心网站,面向社会公开算法运作原理,公开的内容多以商业算法为主,与犯罪治理算法相关的包括“谣言防火墙”的处理流程等内容。百度在“搜索算法规范详解”中梳理并更新了保障搜索安全与内容质量的搜索算法,公开了针对网页内容质量问题的飓风算法 3.0、针对 B2B 领域低质问题的细雨算法 2.0、针对恶意劫持行为的烽火算法 2.0 等。^{⑤④}

可见,国内头部互联网企业为响应监管的要求,普遍履行了算法公开及解释说明义务,起到了良好的示范效应。这种公开属于现象描述层的有限公开和可接受解释,其公开及解释方式易于社会公众理解,保障了公众的知情和监督权利。现象层的有限公开实际相当于算法治理的网络公告,不仅不涉及商业秘密,更无从被不法分子利用而规避算法,且对不法分子形成了一般性的威慑效应。未来,互联网平台应进一步扩大所部署和使用的犯罪治理算法的公开规模,创造敏捷高效的外部监督条件。

2. 公权力机关的算法公开

“警务活动本身具有极强的封闭性,算法也表现出一定的黑箱性,大数据警务呈现出双重封闭特征。”^{⑤⑤}公权力机关的算法使用涉及一套不同于纯粹私主体算法使用的法律框架,因而阻碍了技术和操作细节的透明度实现。^{⑤⑥}相对于互联网平台,公权力机关的算法治理更具公开必要性。域外公权力机关的犯罪治理算法公开主要包括三种类型:

其一,政府集中公开。2022 年,荷兰确立算法登记的强制性要求,建设“公共算法登记系统”(Public AI Registers)。截至 2026 年 2 月 2 日,该系统共公开了包括防控毒品犯罪和暴力犯罪的“Top600”算法、针对青少年高危人群的“安全替代方案”、基于犯罪热点地图的犯罪预测系统、反洗钱监测系统等犯罪治理算法在内的 1343 个算法系统。^{⑤⑦}

^{⑤③} 《变更支付宝隐私权政策的公告》,载支付宝服务大厅, https://help.alipay.com/lab/help_detail.htm?help_id=201602900850&flag=0, 2026 年 1 月 8 日访问。

^{⑤④} 参见《百度搜索算法规范详解》,载百度搜索资源平台, <https://zy.baidu.com/act/guide?isResponsible=1>, 2026 年 1 月 2 日访问。

^{⑤⑤} 李金锋、陈如超:《大数据警务风险及其规制》,载《中国人民公安大学学报(社会科学版)》2023 年第 5 期,第 135 页。

^{⑤⑥} See David F. Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, Yale Journal on Regulation, Vol.37:800, p.841 (2020).

^{⑤⑦} See *The Algorithm Register of the Dutch Government*, at <https://algoritmes.overheid.nl/en> (Last visited on February 2, 2026).

2024年,英国要求相关部门将公共算法公布于“算法透明度记录标准”(Algorithmic Transparency Recording Standard)之中,公开的信息包括算法描述、所有者和责任、决策流程和监督机制、技术工具和数据、影响评估结果、算法风险及应对措施等。该系统有两项与犯罪治理有关的功能,即汉普郡和泰晤士河谷警方使用的“家庭虐待风险评估工具”和西米德兰兹郡警方使用的“性犯罪的探索性分析”。^{⑤⑧}其二,警务部门公开。美国芝加哥警察局在官网公开“犯罪与被害风险模型”,包括模型介绍、模型所使用的信息、风险预警原理、模型准确性以及带有具体算法程序的技术描述等。^{⑤⑨}其三,开发者公开。2019年4月,Equivant公司在官网发布《COMPAS Core 使用者指南》,公开了COMPAS评估解读方法、风险量表的基本信息、信度和效度验证结果等内容,^{⑥⑩}保障了再犯风险评估系统的透明度。

同时,域外国家的公开豁免及部分加密制度亦同步推进。英国2024年发布《“算法透明度记录标准”强制范围和豁免政策》以专章规定算法公开豁免,指出在算法面临被不法分子利用的风险、国家安全风险、侵犯知识产权或泄露商业秘密风险等情况下,国家机构适当的应对措施在于剔除相关细节信息后的有限公开。^{⑥⑪}欧盟《人工智能法案》引言第131条也规定:“对于执法、移民、庇护和边境管制领域的高风险人工智能系统,登记义务应在数据库的安全非公开部分履行。对该数据库安全非公开部分的访问,应严格限制于欧盟委员会以及仅能查看其本国部分的市场监管机构范围内。”

遗憾的是,我国尚未形成一套针对公共算法的法律制度,更遑论公共算法公开,公权力机关的算法公开仍处于算法规制的空白地带。为消除社会疑虑和增进算法信任,有必要参照域外经验,推动我国的犯罪治理算法公开制度建设。

一方面,以主动公开和依申请公开相结合的方式公开犯罪治理算法。参考域外的政府集中公开、警务部门公开和开发者公开三种类型,依据《个人信息保护法》第17条个人信息处理者的公开告知义务和《政府信息公开条例》第4条政府信息公开工作机构职能之规定,采取“警务部门公开”等算法部署者、使用者公开的路径更契合我国法治体系,即算法公开由公安机关等算法控制者的信息公开机构负责。具言之,在主动公开层面,应探索适应数字政府发展趋势、方便公众获知的公开模式,以“公开文本规则+释明技术规则”等方式公开算法模型的描述性信息,明示部署和应用算法的主体、目的、

^{⑤⑧} See *Algorithmic Transparency Recording Standard - Guidance for Public Sector Bodies*, at <https://www.gov.uk/government/publications/guidance-for-organisations-using-the-algorithmic-transparency-recording-standard/algorithmic-transparency-recording-standard-guidance-for-public-sector-bodies> (Last visited on August 2, 2025).

^{⑤⑨} See *Crime and Victimization Risk Model (CVRM)*, at <https://www.chicagopolice.org/wp-content/uploads/FACT-SHEET-Crime-and-Victimization-Risk-Model-1.pdf> (Last visited on January 12, 2026).

^{⑥⑩} See *Practitioner's Guide to COMPAS Core*, at <https://equivant-supervision.com/wp-content/uploads/2023/05/equivant-Practitioners-Guide-to-COMPAS-Core-MAKE.pdf> (Last visited on February 27, 2026).

^{⑥⑪} See *Algorithmic Transparency Recording Standard (ATRS) Mandatory Scope and Exemptions Policy*, at <https://www.gov.uk/government/publications/algorithmic-transparency-recording-standard-mandatory-scope-and-exemptions-policy/algorithmic-transparency-recording-standard-atrs-mandatory-scope-and-exemptions-policy> (Last visited on October 31, 2025).

方式和范围等,为针对犯罪治理算法的外部监督提供条件,保障算法权力合规运行。在依申请公开层面,应保障此前被边缘化、客体化的个体知情、参与、监督等权利,以此跨越“数字鸿沟”、保障过程公平。算法公开包括“以模型为中心”的解释和“以对象为中心”的解释,前者提供关于算法模型的广泛信息,无涉具体决策;后者追求在算法本体、决策主体、算法对象之间建立一种有意义的解释关系。^②一般个体有权向有关部门申请公开算法模型信息,以保障个体受算法影响时的能动性和可干预性,实现“以模型为中心”的解释;在算法决策引发争议时,作为受算法影响者的个体可向作出决策的部门申请公开算法决策的规则及理由,为嗣后的申诉、救济提供制度渠道,实现“以对象为中心”的解释。

另一方面,引入利益衡量机制,场景化判断特定算法是否属于国家秘密、工作秘密或不宜公开的信息,以此建构犯罪治理算法的公开豁免规则。英国、加拿大等国提出了基于“公共利益测试(the public interest test)”的利益衡量机制,要求政府权衡信息公开所涉公共利益,只有不公开的公共利益(避免妨碍犯罪防控等)超过公开的公共利益(透明度和可问责等)时,才能豁免公开或拒绝确认持有信息,否则必须公开信息。^③基于此,可将利益衡量机制引入公共算法公开豁免规则,促进趋向过程公平的外部规制。首先,如果法律法规未禁止公开,有关部门应结合《保守国家秘密法》《政府信息公开条例》等规定,运用利益衡量机制,基于具体场景判断是否应将某一算法纳入公开豁免范围。判断的原则包括信息必须与合法目的相关、公开将对该合法目的造成实质损害、对合法目的损害必须大于公开所获取的公共利益。^④其次,如果某一算法的全部信息符合公开豁免条件,也应履行备案或登记义务,全周期记录算法运行情况;如果某一算法仅有特定细节或部分内容无法公开,则应将除标注密点之外的信息对社会公开并释明。最后,还应建立豁免范围的动态评估机制,根据情势变化及时调整算法公开的豁免范围。

五、结 语

当下,算法深度嵌入且深刻形塑了犯罪治理体系,极大提升了治理能力,但算法偏差、失范等多重风险日益显现,法律规制的滞后更加剧了算法风险。既有法律多强调算法对治理赋能,但对犯罪治理算法的法律规制尚处于起步阶段,关于算法治理的目的、影响评估、适用范围、必要性理由、数据来源、正当程序、人工监督、风险化解及可解释性等事项的规制条款亟待明确和细化,受算法影响者的权利保障仍待改善。为此,必须探寻

^② See Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking for*, *Duke Law & Technology Review*, Vol.16:18, 55-58 (2017).

^③ See Information Commissioner's Office, *The Public Interest Test*, at <https://ico.org.uk/for-organisations/foi/freedom-of-information-and-environmental-information-regulations/the-public-interest-test/> (Last visited on January 9, 2026).

^④ See ARTICLE 19, *The Public's Right to Know: Principles on Right to Information Legislation*, at https://www.article19.org/data/files/RTI_Principles_Updated_EN.pdf (Last visited on January 13, 2026).

算法正当使用的法律原则和具体规则,将算法对犯罪风险及个体的控制纳入法治轨道。算法公平原则为算法规制体系建构提供了价值立场和分析框架,确立了算法权力运行与国民权利保障相匹配的实质公平要求,并探索出回应性规制和预防性规制的双重路径。回应性规制以算法救济和算法问责为主要机制,强调对结果公平的保障;预防性规制包括算法影响评估和算法公开,聚焦过程公平的实现。

必须承认,算法规制体系建构非一日之功。依托网络法规范相对丰富的制度供给,针对互联网平台运用犯罪治理算法的救济、评估及公开等机制已有一定基础,今后还需进一步完善专门性规制条款。相对而言,针对公权力机关运用犯罪治理算法的规制较薄弱。考虑到回应犯罪治理的严峻挑战以及人工智能应用的探索性特征,算法规制体系的建构将会是一个渐进的过程。这有赖于未来的人工智能立法从基本制度上整体筑基,厘定算法权力的法治边界,增强算法治理的可解释性和可控性,重构犯罪治理中的人机关系;还有赖于公权力机关在算法治理中严格遵守算法运行的合规底线、敏捷回应社会诉求、充分保障公众参与监督的权利,以算法向善增进人民福祉与公共利益。

Abstract: The algorithmic governance of crime has become a fundamental strategy of the Peaceful China Initiative, with applications covering scenarios such as location scoring, stability maintenance risk control, big-data-based crackdown on prostitution, and cybercrime risk control. The crime governance algorithm characterized by ‘monitoring and screening’ has significantly enhanced the ability of crime governance. However, it has also led to problems such as an incomplete legal basis, intrusion into individual lives, and misconduct of algorithmic power. The principle of algorithm fairness with both consequential and procedural dimensions shall be the legitimacy basis of algorithm regulation to promote the transition of algorithmic governance of crime from ‘technology-based governance’ to ‘rule-based governance’. This should follow a dual path of responsive regulation and preventive regulation. Responsive regulation can correct algorithmic errors to prevent the erosion of result fairness through algorithmic relief and accountability. Preventive regulation can prevent deviations of algorithms from process fairness through algorithmic impact assessment and algorithmic transparency. They develop algorithms for good and guarantee that algorithmic governance consistently operates within the framework of the rule of law.

(责任编辑:王 楷)