【比较与借鉴】

大数据视野下犯罪预测的机遇、风险与规制

——以英美德“预测警务”为例

胡 铭 严敏姬

【摘 要】犯罪预测是警察部门进行犯罪预警与侦查的重要手段。大数据背景下的犯罪预测兼有传统犯罪预测原理和大数据分析技术的双重特色。犯罪预测大数据化改变了传统警务的执法模式,提升了犯罪预测的效率与精确度,确保了执法过程证据留痕与可追溯。与此同时,数据化的犯罪预测中存在的“黑数据”现象、数据获取过度侵犯个人隐私、算法不透明导致的歧视偏见以及数据壁垒的存在,给科学的犯罪预测造成一定风险。为应对风险,有必要优化犯罪预测中的数据选择标准,在数据收集时坚持信息“个人自决”原则和比例原则,一定范围内促进算法公开与透明,加强数据之间的交流与共享。

【关 键 词】大数据;犯罪预测;算法黑箱;数据壁垒;个人隐私

【作者简介】胡铭,浙江大学光华法学院常务副院长,教授,博士生导师,法学博士,研究方向:刑事诉讼法学、大数据与人工智能法学;严敏姬,浙江大学光华法学院博士研究生,研究方向:刑事诉讼法学(浙江 杭州 310008)。

【原文出处】《西南民族大学学报》:人文社会科学版(成都),2021.12.84~91

【基金项目】国家社科基金重大项目“深化司法体制改革和现代科技应用相结合的难点与路径研究”(18ZDA137)、浙江大学国家制度研究院成果。

“数据”在刑事司法决策领域的应用日益受到关注。以传统统计学为基础的犯罪预测早在20世纪中叶就已在西方国家的警察部门得到广泛运用。该过程被认为是“精算司法”(actuarial justice)在刑事司法领域兴起的体现。[1]随着信息技术的发展,大数据、云计算和人工智能等技术的发展与普及,一个以海量信息和数据挖掘为特征的大数据时代已经到来。2011年,美国《时代》杂志将“预测警务”称为年度50项发明之一。[2]区别于传统结构化、抽样、假设检验的犯罪预测模式,大数据背景下的犯罪预测侧重于数据的大样本、全样本分析并通过算法寻找相关变量之间的关系。在此基础上,预测警务得以快速发展。[3]近年来,国内公安机关依托大数据平台建设,利用大数据进行犯罪预测、预警的现象亦非常普遍。然而,大数据在犯罪预测领域的应用呈现类似“双刃剑”的局面,可谓是机遇与风险并存。在面对犯罪预测数据化带来的风险时亟需明确一定的规制路径。域外“预测警务”已经有多年的实践,其经验和教训对于我们具有借鉴意义。鉴于此,本文尝试以比较研究方法切入,探讨大数据视野下犯罪预测的机遇、风险与规制。

一、大数据背景下犯罪预测的本质与创新

大数据背景下的“犯罪预测”被西方学者喻为“旧把戏,新技术”(old trick,new tech)。[4]“旧”指的是犯罪预测惯用的理论模型与实践样态与传统背景下基本一致;“新”指的是犯罪预测的样本选择与分析方式在大数据背景下具有“数据化”的海量特色。

(一)本质:犯罪预测的理论模型与实践样态

随着政策科学的发展,以量化分析为基础的政策分析得到极大成长并强调现代科学技术和各种研

究论证方法的使用。[5][P.17]犯罪预测就是依靠可靠性日益提高的数据以及分析技术,作出正确的犯罪预警,达到科学地预防和控制犯罪。然而不论信息技术如何发展,犯罪预测所赖以维系的理论模型和实践样态并没有发生根本性的改变。

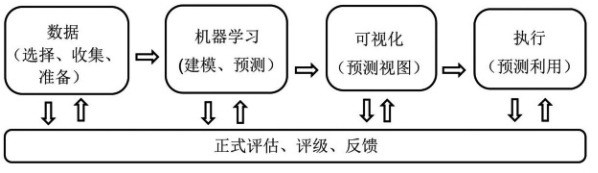
近重复理论(Near Repeat Theory)和风险地形建模(Risk Terrain Modeling)是犯罪预测两个主要的理论模型。近重复理论旨在"识别和解释某些犯罪表现出的在同一地点产生重复犯罪活动的现象"。该理论认为,一旦特定地点发生犯罪,统计学上该地点和附近区域发生犯罪的可能性就增大。在发生首次犯罪后的短时间内,附近环境将可能遭受其他类似的犯罪事件。[6]近重复理论在财产犯罪尤其是入室盗窃案件中显示出极强的近乎重复模式。此时,通过大数据的收集与分析,当某地出现近重复犯罪时,警方就可以加强对特定地域的巡逻,借以威慑犯罪。风险地形建模则更多侧重于社会、物理空间和行为因素间的动态交互作用。风险地形建模的创建首先需给各个因素配值,每个因素形成单独的风险地图层,最后当所有图层在GIS系统中组合在一起时会形成一个风险地形图。风险值越高,代表该位置发生犯罪事件的可能性就越大。[7]风险地形建模不仅可以适用于入室盗窃等案件,还可以有效应用于预防暴力犯罪。随着数据量的增大以及交互式信息技术的进步,风险地形的预测及预警机制正愈加精确化。

根据预测对象不同,犯罪预测的实践样态可分为以人为预测对象和以犯罪区域为预测对象。其中,对犯罪人再犯罪的风险预测是"预测警务"的主要运用场合之一。例如,英国达勒姆郡警察局和剑桥大学合作开发的随机森林(a random forest)预测危害风险评估工具HART(Harm Assessment Risk Tool)。[8]该系统使用达勒姆郡警察局2008年至2012年共104,000个监禁案例,并提取案例中记载的年龄、性别、邮政编码、犯罪历史以及犯罪类型等信息。[9]通过HART模型,能对犯罪者未来24个月的风险进行预测,当犯罪者被捕后,警察就会利用该系统对其进行评估并作出是否羁押的决定。

区域犯罪预测是对一个场所、社区、城市、省或国家的犯罪现象进行预测,评估其犯罪发生的趋势,为区域犯罪预防提供基础。德国Precobs软件是区域预测的典型,该软件主要利用过往犯罪的数据(如位置、时间、事件和其他细节)等查找"高风险"区域。[10]其预测过程可概括为以下步骤:首先,定义检测重复犯罪的标准;其次,计算在逆向分析中已经检测到近重复数据出现的区域并创建空间预测。通过逆向模拟测试标准和计算的区域,以查看所选假设是否有效。当这些区域记录了新的触发要件时,将创建预测(警报),[11]以安排警察的执法活动。

(二)创新:犯罪预测的样本选择与分析方式

大数据的运用,创新了以下过程。"预测警务"的开展过程是大数据公司与警务部门间不断进行数据交换与预测执行、反馈的过程,大致由以下步骤组成:



可以看出,数据是"预测警务"开展的前提与核心。相较于传统犯罪预测方式,大数据背景下的犯罪预测最具"颠覆性"的发展当是样本选择和分析方式的变化。

在信息化平台并未普及时,有关犯罪的信息主要通过纸质方式予以存储。早期运用数据分析进行犯罪预测时,警察需借助纸质地图,通过人工方式将案件有关的各种信息如时间、地点等跟纸质地图结合起来,从而寻找其中的规律。此时,由于人力资源的限制,样本选择非常有限,往往通过抽样方式进行采集,预测结果会有较大偏差。抽样调查中,无论样本选择多大,都会存在一定程度争议。相较而言,大数据背景下的犯罪预测采用的是全样本分析。由于数据处理技术增强,通过机器学习可以同时处理数以百万计的数据,偏差大大降低。此时数据样本选择也不再局限于以往发生的真实案件,社交媒体上

的数据也成为犯罪预测的主要来源。大数据技术所具有的实时搜集、处理数据的能力凸显。

同时，传统犯罪预测主要依靠经验分析、因果关系分析，而大数据犯罪预测更加注重数据之间的相关性分析。具体而言，大数据分析将“抽样”转变为“普查”，内含的数据算法将案件事实间的联系从因果关系证明转向相关性证明，根本上改变了社会科学实践方式。传统犯罪预测中，囿于技术局限，犯罪预测所依赖的数据是传统的社会科学数据，包括真实发生的案例、实验、调查和访谈等。数据的结构化、滞后性导致此时的犯罪预测体现为一种循证式、被动式的分析。[12]具体预测离不开警察在办案过程中所积累的经验，因而这种预测是初级的、不成熟的。大数据技术所具有的相关性分析可以减少人们处理数据时主观假设的影响，完全依靠数据之间的相关性进行阐述。此外，社交网络的发展使得当今社会的数据来源更为多样。数据的海量化、多样化、非结构化是信息社会区别于传统社会的一大特征。此时传统分析方式难以处理此种数据。海量的、非结构化的数据催生了开发用于分析此类数据的软件算法研究。[13]大数据分析依赖的文本挖掘、数据挖掘以及机器学习功能使得实时分析数据成为可能。此时的犯罪预测体现为一种相关性、主动式的分析方式。准备数据、建立模型、机器学习、预测可视化、执行成为大数据犯罪预测必不可少的环节。[14]

二、犯罪预测数据化的新机遇

犯罪预测是犯罪预防必不可少的前提条件。在大数据背景下，犯罪预测的手段和方式更加智能化，各种可视化技术和机器学习算法被运用到犯罪预测中，从而为犯罪预防带来了非常重要的新机遇。

(一)改变传统警务执法模式

传统“标准警务”往往体现为事后的应对。在有限的警力、经费限制下，警务资源更多地投入到案件侦破、打击现行犯罪中，从而形成热点聚焦和大案主导的反应式警务模式。然而，反应式警务模式正面临边际效能困境，在一定时期、一定区域内更多的警力资源投入对于整体警务效能提升的作用呈递减态势。在美国，1980年代起警察管理部门开始把警务资源从巡逻转向处理公民的报警电话，因此强化了孤立的应对性警务。然而，实践证明传统的事后应对、反应式警务只能促进打击犯罪，并不能对犯罪预防起到非常大的积极作用，因而也难以达到有效控制犯罪的目标。

在人员不足、经费有限的情况下，如何合理配置警务资源，提升警务效能成为现代警务改革的重要问题。信息科技的进步为此提供了解决之道。通过加强警务系统的信息化、数据化水平，促使警察执法模式从传统“标准警务”“反应式警务”向“智慧警务”“预测警务”过渡。基于大数据的运用，犯罪预测的数据作用凸显。警察部门根据犯罪预测所形成的可视化视图，可以更加合理安排日常工作。在2009年美国“预测警务”研讨会上，旧金山警察局局长乔治·加斯科恩曾表示，“有了预测警务，我们可以在适当的时间把警察放到适当的位置或提供其他服务来打击犯罪，并且可以使用更少的预算做到这一点”。[15]也因此，警察巡逻的随机性递减，警务执法模式由传统走向数据化、智能化。在我国，越来越多的公安指挥中心也从单纯的接警、派警，变成集数据、情报、指挥、服务于一体的综合平台，从而更好地预知预警、防控风险。

(二)提升犯罪预测的效率与精确度

在传统警务模式中，犯罪预防往往采用随机预防式巡逻。有观点认为，如果警察开着有巡逻标志的警车定期在小区巡逻，即使他们没有特定目标，也会震慑住潜在的犯罪嫌疑人。20世纪70年代，华盛顿的警察基金会在密苏里州的堪萨斯市做了一个实验，结果显示，随机预防式巡逻对犯罪率没有产生实质影响。[16](P.48-49)与此相反，大数据在犯罪预测中最直接的作用就是提升犯罪预测的效率与精确度，从而达到更好的犯罪预防效果。

首先，犯罪预测数据化可以提高犯罪预测的效率。在传统的警务模式中，警察巡逻是随机的，此种无目的出警是对警力资源的浪费。在大数据环境中，通过数据预测，警察执法变得更为“智慧”。预测

警务理论认为,在统计上更有可能犯罪的地区应该有针对性地增加警力。[17]通过将随机巡逻变为“定点巡逻”,通过数据指导警察巡逻的模式,甚至是特定的时间、日期和地点,稀缺的警察资源可以集中在犯罪风险较高的地区,促进警察资源的有效利用。自2017年来,德国黑森州国家刑事调查局通过KLB-operativ内部开发软件,使辖区内每个警察都可通过智能手机的应用程序对入室盗窃案件进行预测。该应用程序每天早上更新,以映射过去十天来有价值的入室盗窃案并突出显示每日热点地区。[18]

其次,犯罪预测数据化可以提高预测的精确度。犯罪预测某种程度上是犯罪事件是否发生的概率问题。在传统的犯罪预测中,犯罪预测的精确度并不高,因而针对性犯罪预防的效果并不理想。在数据体量及质量不高的情况下,警察所能获取的预测结果相对有限。随着信息技术的发展,预测数据的海量化以及精准化,犯罪预测的精确度可以得到相当程度的提高。一方面,数据所具有的客观性,可以弥补人类感知的脆弱性,从而增强预测结果的客观性与精确度。美国纽约警察局曾被曝种族歧视严重。该警察局超过95%的情报调查以穆斯林主体为目标,还曾被爆出监视黑人生活。[19]除种族外,犯罪嫌疑人的性别、阶级、财富等都可能对警察的主观预测产生影响。然而在大数据世界中,此类个性化数据在某些情况下可予以一定限制,避免因警察的主观预测而对犯罪预测结果产生影响;另一方面,预测结果的精确性又可以缓和警察执法过程可能造成的社会恐慌。在预测结果产生后,警察部门需要对预测结果进行反应。它通常表现为警察到一些热点地区巡逻或者定点逮捕犯罪嫌疑人。如果预测不够精准,很可能会打草惊蛇,甚至引起周围群众的恐慌心理。

(三)确保执法过程证据留痕与可追溯

警察部门作为行政执法机关,其执法行为依据的方式、方法、过程都可能面临相对人和社会公众的质疑与监督。在大数据犯罪预测过程中,犯罪预测软件是警察执法的工具,犯罪预测结果是警察部门采取相应手段的前提。若警察部门根据预测结果采取了一定行为,后续就可能面临需要对执法行为进行说明甚至产生责任承担问题。此时,数据化的犯罪预测通过数据留痕可以使犯罪预测的过程通过可视化的方式保存下来,进而确保执法过程的证据留痕以及后续的责任承担问题。[20]

通常情况下,警察采取行动前的准备工作很难被完整记录。数据化的犯罪预测通过预测软件则可将这一过程自动记录下来。通过记录,警察可以说明他们访问了哪些数据库,使用哪些步骤、条件进行查询,从而证明他们在调查潜在犯罪嫌疑人时所采取的步骤。例如,在警察采取行动逮捕犯罪嫌疑人后可以说明,他已经访问了相关个人信息数据库并结合车牌进行检查,并用这些信息佐证他的怀疑。通过数据留痕、数据库的访问记录等可视化方式进行说明不仅可以体现警察执法行动的依据,甚至可以简化法官对警察合理怀疑的判定。[21]此外,犯罪预测数据化还可以在警察部门内部形成一种进行数据审核、记录收集标准的良好风气。通过记录,相关部门可以随时检查警察执法的依据,了解哪些因素会对警察逮捕犯罪嫌疑人产生影响,并将此作为一种内部监督策略用于后续的问责机制中。[22]例如,在警察通过犯罪预测当场抓获犯罪嫌疑人并予以逮捕的情况下,如果逮捕错误,在后续的内部追责过程中,可令警察对预测过程进行说明。

(四)促进刑事侦查理性与经验的平衡

依托大数据之犯罪预测是连接传统侦查向信息化的大数据侦查转型的纽带。大数据侦查的核心就在于利用大数据技术进行犯罪预测与打击。[23]传统侦查决策讲求经验决策,主要凭借决策者在侦查工作中积累的办案经验或形成的办案直觉来指导办案。然而,此种依赖于侦查人员个人主观经验与判断的侦查模式在大数据时代显得捉襟见肘。大数据时代的侦查工作亟需理性主义及精算司法的普及。然而,理性主义和经验主义的极端化都不可取。西方学者曾提出“大数据经验主义”概念,认为大数据时代“理论终结”。[24]如若延伸到刑事侦查领域,是

否意味着大数据时代的刑事侦查只需要数据理性而不用讲求传统经验、理论分析的运用？其实大数据时代侦查工作的开展应讲求理性主义与经验主义的平衡。在犯罪学领域，研究人员早已将社交媒体上的数据作为研究数据的来源。作为用户自我生产的数据，研究人员可以看到用户实时生成的自我报告数据。这些数据可以补充或替代传统来自实验、调查和访谈等途径的社会科学数据。[25]此时数据的海量与混乱导致只有使用全体数据，而不是样本数据才能对相关问题进行解释。犯罪预测背景下的大数据侦查即是如此。在面对新型犯罪、网络犯罪等特殊犯罪类型时，大数据侦查尤显必要。相对而言，面对传统类型的犯罪案件，侦查人员个人的主观能动性以及经验决策在案件侦查过程中仍具有重要地位。

三、犯罪预测数据化的可能风险

随着信息化的发展，计算机处理器的加快以及存储数据量的增多，先前离散的数据网络得到新的发展。[26]基于数据质量和数量对“预测警务”的重要性，导致警察部门和相关合作公司会积极寻求大量的数据应用于相关预测系统。由于缺乏明确的监督和制约法规，在数据的获取、运用、整合阶段都易产生相应的风险。

(一)数据搜集——“黑数据”现象导致歧视

黑数据(black data)亦称为坏数据(bad data)，是所有警察部门在大数据预测时所面临的共同难题。算法是一个黑箱，进入黑箱的数据对其结果的产生起决定作用。当前，预测警务所采用的数据质量并没有达到理想状态。一方面，面对庞大、不断增长的网络数据系统，缺乏相应共享数据的质量控制。警察部门可以在各种网络社交媒体上搜集数据，而此类平台的数据往往缺乏监管，可靠性存疑；另一方面，数据来自现实世界，带有种族差异和不公平的天性。“警察部门一直在大量收集数据，但我们不知道数据来源是否可靠、有效和干净。因而，我们需要对数据收集进行监督以确保数据干净。”[27]若缺乏数据搜集过程中的质量监督，预测结果可能会导致偏见的产生，甚至扩大现有偏见。

刑事司法制度有其自身固有的偏见。在西方刑事司法系统中，种族和宗教容易引发歧视。经济、社会地位上的弱势群体以及一些少数族裔人群，受到暴力以及其他犯罪侵害的概率明显更高。进而有德国学者指出，“预测警务”可能会放大现有的偏见和歧视。例如，警察在被定义为“热点”的地区更频繁地巡逻，在该地区将记录更多的犯罪报告，同时，在未来预测中该地区的犯罪数量占比将更重。[28]美国Palantir和DAS预测系统的反对者认为，警察依据自己的种族偏见使用数据监控，从而监控甚至逮捕更多的人，是“种族主义的循环”。[29]2016年，美国非营利组织人权数据分析小组(Human Rights Data Analysis Group)利用Predpol算法，在奥克兰市地图上推演了由警察记录数据所构成的毒品案件逮捕的空间分布。结果显示，逮捕行动主要集中在非白人和低收入人口占主导的社区。如果将这些数据输入机器学习算法，则可能加剧人口逮捕的不平等分布。[30]因而，在数据搜集阶段如何最大程度避免黑数据，提高搜集数据的质量至关重要。

(二)数据获取——过度侵犯个人隐私

数据获取是“预测警务”开展的前提。在数据获取过程中，“个性化数据”的过度整合与利用是“预测警务”过度侵犯隐私权的重要体现。美国洛杉矶警察局曾与Palantir公司合作整合个人数据用于警务监测和预防。Palantir平台可以获取自动车牌读取器(automatic license plate readers)的数据以搜集所有公民的信息，而不限于犯罪嫌疑人。随后，系统可以绘制数据地图以帮助警察追踪城市中的车辆和人员，使执法部门了解驾驶员的典型出行方式并识别差别。据此，警察部门可建立一个庞大的个人信息数据库，即使从未与警察直接接触，个人信息亦可能被存储在数百万个数据点中，包括驾驶汽车的型号、家庭住址等。[31]美国纽约警察局与微软公司曾合作开发一项大数据犯罪预防和反恐技术——DAS系统。DAS可以利用摄像机、车牌读取器和射频感应器创建纽约市的实时监测地图。该系统与整个纽约市的

私人闭路电视监控(privately-owned CCTV cameras)合作,并与多个非纽约警察局的情报数据库进行对比。在DAS运行后,纽约市市民认为其严重侵犯了个人隐私权以及免于无根据监视的权利。2018年,纽约市议员凡妮莎L.吉布森(Vanessa L. Gibson)提出了《监视技术公共监督法》(The Public Oversight of Surveillance Technology)以期对此进行规制。同时,美国各地开始制定协议,确保不会滥用自动车牌阅读器和其他监视技术。[32]

(三)算法黑箱——预测缺乏公正、透明

在大数据犯罪预测系统制定过程中,往往需要计算机科学家的帮助,因为绝大部分警务人员没有编写代码的能力。除数据分析人员外,几乎所有的数据系统对用户来说都是"黑匣子"。即使是简单的数据库,用户也无法理解。数据是算法运算的提前。然而,数据获取阶段存在"黑数据"现象,来自现实世界的数据带有种族差异和不公平的天性,同时新的数据技术具有一定的保密性,加之数据中的隐私和宪法保护不够,[33]导致公众本身对使用这些数据的算法带有天然的不安全感。同时,商业领域保密措施的存在意味着大数据算法等私营供应商可以规避公共部门的透明度要求,使大数据警务变得难以监管和规范。在缺乏透明度和问责制的情况下,若法院、公民无法理解该技术,而律师、新闻工作者和学者也无法质询该数据,那么谁能信任算法产生的结果?[34]

此外,当以大数据为基础的算法软件具备机器学习能力时,将使预测过程变得更难掌控。机器学习取决于数据,可以访问的数据越多,学习的效果就越好。数据的质量、数据输入系统的方式以及如何"训练"系统以分析数据可能会严重影响由算法生成的信息的有效性、准确性和实用性。[35]机器学习的结果意味着,预测过程可能会超出其初始编码并使用新的数据产生结果。在这种情况下,使用者甚至程序员也可能不知道算法运行的过程究竟如何。[36]不透明问题带来的影响可能是致命的。有学者认为,在犯罪学语境下,使用增强型机器学习(reinforced machine learning),即机器试图建立为特定问题提供正确答案的规则是一个解决办法。然而,这意味着除非每个犯罪都被举报,警察平等追究所有人犯下的所有类型犯罪,否则不可能有一个能够预测犯罪本身的强化学习系统。事实上,行为会影响结果,导致得到的预测反馈非常有限。[37]因此,算法黑箱实际上可能使社会公众处于"黑暗"之中,我们并不知道什么时候、出于什么原因,我们可能就会出现在犯罪预测系统当中。

(四)数据隔离——信息孤岛的产生

信息孤岛是指相互之间在功能上不关联互动,信息不共享、不互换的现象。每当数据系统不兼容或未与其他数据系统集成时,就会发生信息孤岛。信息的不对等、不对称是警务部门开展工作的一大障碍。在英美德三国,不同辖区的警察部门系统独立,使用的数据库也大不相同。就"预测警务"而言,信息数据缺乏共享、联动将会导致警务部门资源利用低下,造成人力、物力、财力的浪费。

然而信息孤岛现象的产生、各个辖区警务部门采用不同数据、不同预测软件具有一定现实依据(此处仅针对财产犯罪而言)。美国犯罪学家认为,财产犯罪是可预见的行为且往往只需该地区有警察驻守就可以制止,但暴力犯罪往往更难预测和制止。因而目前"预测警务"较多适用于财产犯罪。[38]财产犯罪如盗窃等往往具有很强的地域性特征。不同的警察辖区对此规定不尽相同,因而搜集的数据也大不一样。这种数据收集的先天性缺陷导致某一辖区所采用的预测方式在另一辖区并不能适用。以英国HART预测软件为例,该系统收集的数据主要是达勒姆郡警察局的羁押案件。这意味着该系统的适用范围具有局限性,不能在其他地区警局得到应用。此现象在德国警察部门体现得更为明显。德国联邦各州警察部门针对"预测警务"开发了相应专属的软件系统。[39]这些系统同HART一样,都限于某辖区使用。相较而言,美国PredPol、Palantir等数据平台在一定程度上实现了利用海量数据进行数据整合的优势,适用面较广,因而也得到美国很多警察部门的采

用。当前,预测警务数据孤岛现象更为严重的问题在于:一些本可以整合、共享的数据库彼此之间并没有实现有效的数据共享与流通,从而造成资源的浪费。

四、犯罪预测数据化的规制路径

大数据视野下的犯罪预测机遇与挑战并存。西方国家预测警务面临的风险亦是我国公安机关在实践过程中正在面临的问题。通过一定举措对这些风险予以规制是各国预测警务发展的必然要求,也是值得我国公安机关学习之处。

(一)优化数据选择标准

大数据支持下的预测警务,数据是预测的灵魂所在。数据搜集阶段存在黑数据,容易对犯罪预测产生不良影响。因而需要对所选数据进行一定限制,避免无关因素影响到预测结果。优化数据选择是为了后期算法利用数据处理时最大程度确保结果的真实性与客观性,因而在数据收集选择阶段就需要把数据的真实性和客观性作为数据选择的标准。

首先,确保数据的真实性。目前,警察部门获取数据的来源除自身所做的犯罪记录以及公共部门的犯罪监测数据及社会性数据外,有很大一部分来自第三方私人平台。在涉及警察等公共部门提供的数据时,为了预测结果的真实性需要尽可能收集与犯罪有关的各方面信息,减少犯罪数据不足所带来的偏差,扩大数据收集范围;在涉及第三方私人平台的数据时,需要制定相应的数据过滤机制,剔除不真实的数据。其次,确保数据的客观性。保持数据客观性是为了避免偏见与歧视。目前,美国一些司法辖区都进行了一定程度的数据收集和审查改革,但是并没有相关法律对警察执法期间形成的非法和有偏见的数据进行限制。[40]因而,未来需要改革相关法律,限制或禁止警察使用非法和有偏见的数据,避免由于反馈回路的存在对未来的数据处理与预测的作出产生不良影响。再次,数据收集、选择阶段要完全做到没有黑数据的存在并不现实。要想真正解决这一问题,还得通过明确的归责制度来解决。预测系统的存在具有一定证据留痕的效果。警察行为的作出是否合法、合理可以通过回溯相关数据库与警察作出决定时的衡量因素来判断。由警察部门承担在采用预测系统时使用的数据是真实客观、无偏私的说明责任,可以在一定程度上化解该问题,并可以通过事后的追责程序,追究警察预测数据的疏忽或滥用导致对犯罪嫌疑人甚至是普通公民权益产生侵害的问题。

(二)坚持“个人自决”与比例原则

大数据使隐私侵权变得十分容易、普遍,甚至有人预言:“大数据时代,隐私权已死”[41]。数据挖掘使民众隐私权遭受前所未有的侵犯。“互联网是有记忆的”,民众并不能寄希望于一切侵犯个人隐私对自身造成伤害的行为会随着时间流逝而消失。因而,最根本的解决办法在于从源头保护好个人隐私。

首先,坚持“个人自决”原则。在德国,根据宪法规定“收集和处理个人数据需受到限制”。信息自决权被视为一般人格权的体现并在1983年被德国联邦宪法法院确认为基本权利。《欧洲人权公约》第8条第1款即“人人享有使自己的私人和家庭生活、家庭和通信得到尊重的权利”也是“个人信息自决”的体现。其次,在坚持信息“个人自决”前提下,遵循比例原则。德国联邦法院强调,出于客观确定和有效的理由,只有在应对危险行为时,类似“预测警务”系统的使用才基本被允许。[42]这也是《欧洲人权公约》第8条第2款规定所倡导的。在2009年美国国家司法研究所与司法援助局和洛杉矶警察局合作举行的会议上,司法部司法援助局高级政策顾问托马斯·奥莱利(Thomas O´Reilly)认为:“预测警务的开展不应秘密进行,我们应该一开始就邀请隐私权倡导者和社区领袖来解释该计划,并征求他们的想法和意见,减轻他们的担忧”[43]。在我国,个人隐私保护的力度也在不断增强。随着《网络安全法》《数据安全法》和《个人信息保护法》陆续颁布和实施,个人信息保护达到前所未有的高度。公安机关作为行政部门,在利用职权搜集数据、实施犯罪预测时也必须要保护个人隐私,规范数据使用。

(三)加强算法预测透明度

算法不透明是社会公众对预测警务产生不信任感的重要原因。预测警务的应用必须尝试解决不透明性问题。解决的方法可以包括加强公众对算法决策的了解,明确警察部门在搜集数据时的责任承担问题以及在相对范围内披露算法预测的过程等。

首先,应让公众了解大数据监管与日常生活中的其他算法决策并无不同。预测警务的开展并不针对特定的个人,而只是通过一系列的数据、算法预测何时何地犯罪可能发生的概率,借以合理安排警察部门的日常工作。同时,应对透明性问题并不意味着警方需要提供更多的信息,而是需要明确一定的责任承担问题。公众需要的并不一定是公布算法决策的具体过程,而是使用该算法的原因,其中的衡量因素是否公正、是否包含偏见等。民众对透明性问题的愤怒主要来自政府监视的秘密性质、无限制地进行数据挖掘,而不是实际的技术监视能力。

其次,算法披露只能在一定范围内,要求严格的算法透明是不现实以及不必要的。一方面,在某些情况下,可以允许披露算法,揭示预测过程,以此增强警察执法的可信度。另一方面,商业模式决定了算法的专有技术保密。披露源代码意味着揭示公司在业务上的竞争优势。若是对算法毫无限制地披露,会对相关公司商业利益造成严重的损害。并且,在人工智能中,由于机器具有反馈回路(feedback loops)拥有再学习的能力,机器学习模型每次分析都会有所不同。即使具有技术能力,也可能无法看到基本公式。因而怎么披露算法、向谁披露算法成为亟待解决的问题。有立法者提出:设立一个监管机构或审计部门专门处理算法出现问题时的审计,确保预测过程公平公正。[44]例如,警察根据大数据预测作出相应执法行为后,若产生相应后果,可能对犯罪嫌疑人或普通民众造成不公正的,可以将预测算法披露给中立的监管机构或审计部门,由特定机构进行审查认定算法决策是否公正、合理。

(四)完善数据交流互通

警察部门进行犯罪预测的数据大致可以分为公共主体和私人主体两个来源。从公共主体来看,数据一方面可能是警察部门前期侦查犯罪所保留的数据,例如犯罪嫌疑人、犯罪时间、犯罪地点、犯罪类型等方面数据;另一方面可能是公共监管机构如交通部门所掌握的公民驾驶证信息、车牌号、车型等信息。从私人主体来看,数据可能是通过大数据挖掘技术以及信息搜集技术在社交平台上用户注册时所填写的个人信息以及用户使用相应平台时留下的数据。因而,要形成一个海量的数据共享平台,必须在保障相应隐私的前提下促进数据之间的交流互通。

首先,警察部门需要加强自身的基础数据平台建设。例如,对于先前犯罪的各项数据,要统一数据采集、存储、整理、传输、保存等各个环节的标准,并在最大程度上收集犯罪数据,扩大数据范围。同时,由于不同犯罪所体现出的犯罪特征具有不一致性,需要根据不同的犯罪种类创建相应的数据库。其次,警察部门要加强对交通部门、民政部门等其他政府部门的数据采集和共享,推动政府公共部门之间的数据协作。再次,警察部门要联合私人主体平台,加强对社会范围内的数据收集与共享,扩展数据的规模。最后,不同辖区的警察部门之间要加强内部数据共享,打破地域之间的数据壁垒,从而为预测警务的开展提供海量化的数据资源。

结语

随着数据收集、整合和挖掘技术的进步,以数据驱动为主的技术方法成为警察部门执法的重要手段。同时,越来越多复杂的社会、经济和政治问题需要通过数据进行评估和解决,[45]"数据治理"正成为信息化时代的重要特征。目前,以数据为基础的犯罪预测在我国的运用日益广泛,预测警务这一新的警务运行模式随着国家"大数据战略"的推进也正在实践中深化运用。与域外国家类似,我国预测警务的开展也面临着算法不透明、数据壁垒以及数据获取与个人隐私保护之间的固有矛盾等问题。抓住机遇并直面问题,以法治的进路规范科学的数据预测,公安部门才能更好地预防犯罪的发生,在信息化时代担负起维护国家安全的重要使命。

参考文献：

[1]Sarah Brayne. Big Data Surveillance: the Case of Policing [J]. American Sociological Review, 2017, Vol. 82(5), 977-1008.

[2]Carsten Momsen & China Rennert. Big Data-Based Predictive Policing and the Changing Nature of Criminal Justice[EB/OL].https://kripoz.de/wp-content/uploads/2020/05/momsen-rennert-big-data-based-predictive-policing-and-the-changingnature-of-criminal-justice.pdf.2021-01-15.

[3]Elizabeth E. Joh. Policing by Numbers: Big Data and The Fourth Amendment[J]. Washington Law Review, 2014, Vol. 89, 3568.

[4]Alex Reshanov. How Bias Sneaks into Big-Data Policing [EB/OL]. https://lifeandletters.la.utexas.edu/2020/10/how- bias-sneaks-into-big-data-policing/.2021-01-15.

[5]刘建宏.犯罪干预与预防评估系统回顾研究[M].北京：人民出版社，2015.

[6]Andrew Guthrie Ferguson. Predictive Policing and Reasonable Suspicion[J]. Emory Law Journal, 2012, Vol. 62, Issue 2, 259-325.

[7]Leslie W. Kennedy et al., Risk Clusters, Hotspots, and Spatial Intelligence: Risk Terrain Modeling as an Algorithm for Police Resource Allocation Strategies[J]. Journal of Quantitative Criminology, 2011, Vol. 27, 339-362.

[8]Sheena Urwin. Algorithmic forecasting of offender dangerousness for police custody officers: an assessment of accuracy for the Durham Constabulary model[D]. Cambridge University, 2016.

[9]John Morison and Adam Harkens. Re-engineering justice? Robot judges, computerized courts and (semi) automated legal decision-making[J]. Legal Studies, 2019, Vol. 39, Issue 4, 618-635.

[10]Crime Predicting Computers[EB/OL]. https://www.ebuyer.com/blog/2014/12/crime-predicting-computers/.202101-15.

[11]Seidensticker Kai, Bode Felix & Stoffel Florian. Predictive Policing in Germany[EB/OL]. https://www.researchgate.net/publication/332170526_Predictive_Policing_in_Germany.2021-01-15.

[12]Sarah Brayne and Angéle Christin. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts[J]. Social Problems, 2020, https://doi.org/10.1093/socpro/spaa004.

[13]Janet Chan & Lyria Bennett Moses. Is Big Data Challenging Criminology?[J].Theoretical Criminology, 2016, Vol. 20(1), 2139.

[14]Alkesh Bharati & Dr Sarvanguru RA. K. Crime Prediction and Analysis Using Machine Learning[J]. International Research Journal of Engineering and Technology, 2018, Vol. 5, 1037-1042.

[15]Beth Pearsall. Predictive Policing: The Future of Law Enforcement?[J]. NIJ Journal, 2010, Issue 266, 16-19.

[16][加]欧文·沃勒.智慧的犯罪控制[M].吕岩译.北京：中国法制出版社，2018.

[17]Anthony A. Braga et al., The Relevance of Micro Places to Citywide Robbery Trends: A Longitudinal Analysis of Robbery Incidents at Street Corners and Block Faces in Boston[J]. Journal of Research in Crime and Delinquency, 2011, Vol. 48, 7-32.

[18]Seidensticker Kai, Bode Felix & Stoffel Florian. Predictive Policing in Germany[EB/OL]. https://www.researchgate.net/publication/332170526_Predictive_Policing_in_Germany.2021-01-15.

[19]Ayyan Zubair. Domain Awareness System[EB/OL]. https://static1.squarespace.com/static/5c1bfc7eee175995a4ceb638/t/5f170be2dc09615b852699d7/1595345890732/Domain%2BAwareness.pdf.2021-01-15.

[20]胡铭.电子数据在刑事证据体系中的定位与审查判断规则[J].法学研究，2019(2).

[21]Andrew Guthrie Ferguson. Big Data and Predictive Reasonable Suspicion[J].University of Pennsylvania Law Review, 2015, Vol. 163, 327-410.

[22]ABA Standards for Criminal Justice: Law Enforcement Access to Third Party Records[M]. American Bar Association, 2013.

[23]胡铭，龚中航.大数据的基本定位与法律规制[J].浙江社会科学，2019(12).

[24]齐磊磊.大数据经验主义——如何看待理论、因果与规律[J].哲学动态，2015(7).

[25]Janet Chan & Lyria Bennett Moses. Is Big Data Challenging Criminology?[J]. Theoretical Criminology, 2016, Vol. 20(1), 2139.

[26]Andrew Guthrie Ferguson. Big Data and Predictive Reasonable Suspicion[J]. University of Pennsylvania Law Review,

2015, Vol. 163, 327-410.

[27]Andrew Guthrie Ferguson. The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement[M]. New York University Press, 2017.

[28]Carsten Momsen & China Rennert. Big Data-Based Predictive Policing and the Changing Nature of Criminal Justice[EB/OL]. https://kripoz.de/wp-content/uploads/2020/05/momsenrennert-big-data-based-predictive-policing-and-the-changing-nature-of-criminal-justice.pdf/.2021-01-15.

[29]The LAPD Has A New Surveillance Formula, Powered by Palantir[EB/OL]. https://theappeal.org/the-lapd-has-a-new-surveillance-formula-powered-by-palantir-1e277a95762a/.2021-01-15.

[30]Kristian Lum & William Isaac. To Predict and Serve?[J]. Significance Magazine, 2016, Vol. 13, 14-19.

[31]Sarah Brayne. Op-Ed: One way to shrink the LAPD's budget: Cut costly and invasive big-data policing[EB/OL]. https://news.yahoo.com/op-ed-one-way-shrink-100529937.html.2021-01-15.

[32]Ayyan Zubair. Domain Awareness System[EB/OL]. https://static1.squarespace.com/static/5c1bfc7eee175995a4ceb638/t/5f170be2dc09615b852699d7/1595345890732/Domain%2BAwareness.pdf.2021-01-15.

[33]Andrew Guthrie Ferguson. The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement[M]. New York University Press, 2017.

[34]Andrew Guthrie Ferguson. The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement[M]. New York University Press, 2017.

[35]Pedro Domingos. A Few Useful Things to Know About Machine Learning[J]. Communications of the ACM, 2012, Vol. 55, 7887.

[36]Iria Giuffrida. Fredric Lederer & Nicolas Vermerys. A Legal Perspective on the Trials and Tribulations of AI: How Artificial Intelligence, the Internet of Things, Smart Contracts, and Other Technologies Will Affect the Law[J].Case Western Reserve Law Review, 2018, Vol. 68, 747-781.

[37]Caroline Haskins. Academics Confirm Major Predictive Policing Algorithm is Fundamentally Flawed[EB/OL]. https://www.vice.com/zh-CN/article/xwbag4/academics-confirm-major-predictive-policing-algorithm-is-fundamentally-flawed.2021-01-15.

[38]Police Executive Research Forum 2014. Future Trends in Policing. https://www.policeforum.org/assets/docs/Free_Online_Documents/Leadership/future% 20trends% 20in% 20policing% 202014.pdf.2021-01-15.

[39]Seidensticker Kai, Bode Felix & Stoffel Florian. Predictive Policing in Germany[EB/OL]. https://www.researchgate.net/publication/332170526_Predictive_Policing_in_Germany.2021-01-15.

[40]Rashida Richardson, Jason M. Schultz & Kate Crawford. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing System, and Justice[J]. New York University Law Review, 2019, Vol. 94, 192-233.

[41]徐明.大数据时代的隐私危机及其侵权法应对[J].中国法学,2017(1).

[42]Carsten Momsen & China Rennert. Big Data-Based Predictive Policing and the Changing Nature of Criminal Justice[EB/OL]. https://kripoz.de/wp-content/uploads/2020/05/momsenrennert-big-data-based-predictive-policing-and-the-changing-nature-of-criminal-justice.pdf.2021-01-15.

[43]Beth Pearsall. Predictive Policing: The Future of Law Enforcement?[J]. NIJ Journal, 2010, Issue 266, 16-19.

[44]Katik Hosangar & Vivian Jair. We Need Transparency in Algorithms, But Too Much Can Backfire[EB/OL]. https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire.2021-01-15.

[45]Rashida Richardson, Jason M. Schultz & Kate Crawford. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing System, and Justice[J]. New York University Law Review, 2019, Vol 94, 192-233.

