Doi:10. 11835/j. issn. 1008-5831. fx. 2020. 12. 001

欢迎按以下格式引用:郑飞,马国洋. 大数据证据适用的三重困境及出路[J]. 重庆大学学报(社会科学版),2022(3):207-218. Doi:10. 11835/j. issn. 1008-5831. fx. 2020. 12. 001.

Citation Format: ZHENG Fei, MA Guoyang. Triple dilemma and solutions for the application of big data evidence[J]. Journal of Chongqing University (Social Science Edition),2022(3):207-218. Doi:10. 11835/ j. issn. 1008-5831. fx. 2020. 12. 001.

大数据证据适用的三重困境及出路

郑 飞[1],马国洋[2]

(1. 北京交通大学 法学院,北京 100044;2. 中国政法大学 司法文明协同创新中心,北京 100088)

摘要:大数据证据是对海量数据进行筛选、汇总、提炼、形成结论并在审判中使用的证据。大数据证据不同于"运用大数据技术分析收集的证据",后者并未对传统证据规则形成明显挑战,而前者将导致大数据证据与传统证据规则产生明显冲突,进而引发大数据证据在法庭适用中的三重困境。第一重困境是大数据证据种类与法定证据种类的不适应,这一困境应通过不同阶段的"三步走"策略逐渐解决。第一阶段,应将大数据证据作为一种鉴定意见;第二阶段,应将大数据证据作为独立的证据种类;第三阶段,应放弃将证据种类作为证据门槛的做法。第二重困境是因可靠性质疑而导致的相关性困境,这一困境产生的原因是大数据的黑箱化运行以及大数据技术的复杂性。对此,最为简单直接的办法便是公开算法的历史准确率。其中,算法历史准确率公布的主体应是算法开发者(或改进者),因为开发大数据算法的一个组成部分便是计算(改进)正在进行的算法的准确性。同时,为了保障算法开发者(或改进者)公布的历史准确率具有可信度,还应由政府部门牵头,依托具有相应专业人才、技术支撑和监管能力的行业自律组织负责算法的监管。除此之外,必要时还应寻求鉴定人、专家辅助人进行解释,使一般人能够理解基于"数据经验"产生的关联,从而进一步对大数据证据的可靠性进行判断。第三重困境是因对隐私权的侵犯和"证据偏在"的影响而导致的可采性困境,该困境应通过构建"原则+制度+技术"的融合规制路径来加以解决。从原则角度出发,大数据证据的应用原则包括数据有限使用原则、数据主体"弱同意"原则和数据甄别原则;从制度的角度出发,一方面,应构建大数据技术风险评估系统,对大数据技术的应用进行风险等级评定;另一方面,应引入大数据技术应用的审查机制,这包括大数据监管机构的审查和司法审查。从技术角度出发,应尝试通过"数据脱敏"等更加先进的技术完善隐私保护机制。除此之外,大数据证据第三重困境的化解还需要通过完善证据开示制度等方法增强诉讼的对抗性。

关键词:大数据;大数据证据;证据适用;事实认定;三重困境

中图分类号:D925. 2 **文献标志码**:A **文章编号**:1008-5831(2022)03-0207-12

基金项目:国家社会科学基金重点项目"大数据侦查的程序控制与证据适用研究"(2019AZD024);北京市社会科学基金青年项目"北京市刑事证据保管制度研究——基于冤假错案防治的视角"(16FXC029)

作者简介:郑飞,北京交通大学法学院,Email:zf591014@ 163. com;马国洋,中国政法大学司法文明协同创新中心。

随着大数据时代的到来,大数据的概念正在逐渐为人们所熟知,而大数据技术和理念也不断在司法活动中得到应用,“大数据+司法”的概念不断深入人心。为此,本文试图对大数据证据在法庭中应用所面临的三重困境进行分析,以期更好地推动大数据证据的应用。

一、大数据证据的内涵

大数据证据是指证据提出者通过对海量数据进行筛选、汇总、提炼、形成结论并在审判中使用的证据[1]。一般而言,大数据证据的运用主要有四种形式:一是将大数据的载体作为证据;二是将大数据等量复制的数据副本作为证据;三是将大数据中的部分数据作为证据;四是将大数据分析结论作为证据[2]。大数据证据是一种全新的证据种类。一方面,其在形式上不同于任何传统的证据种类,不仅在法律规定上无法找到其存在的有效性依据,而且在传统科学证据、概率证据等理论范式的讨论之中也无法找到其对应的进路;另一方面,由于大数据技术对于人类一般经验的超越,大数据证据将对传统证据理论构成根本性挑战。这种挑战导致了大数据证据的适用困境。

大数据证据不同于“运用大数据技术分析收集的证据”。后者主要是将大数据技术用作其他证据获取的前提或铺垫,而不用于最终的法庭举证。换言之,在“运用大数据技术分析收集的证据”中,大数据技术主要发挥的是工具性作用,例如“促进情报资源丰富化、线索发现主动化、案情研判智能化”[3]。但其并未对传统证据规则构成真正挑战。以刑事侦查为例,侦查人员通过大数据分析发现了传统的物证、书证或证人证言后,最终在法庭上所提交的证据并非是大数据的分析结果,而仅是依靠大数据分析所挖掘的其他证据。该类证据属于传统的证据种类,对现有证据规则不会产生实质性冲击。如果将“运用大数据技术分析收集的证据”也认定为大数据证据,则有可能引发一份证据(一个证明目的)两个证据种类的困境。例如,通过大数据分析发现的作案工具本是一种物证,如果因其借助了大数据技术便称为大数据证据,那么该作案工具就具有物证和大数据证据双重身份,而这两种证据种类在证据审查时需要遵守不同规则,这就可能引发证据规则的适用困难。

当然,“运用大数据技术分析收集的证据”与大数据证据也并非完全对立,二者均利用了大数据技术。因此,二者在证据审查时可能面临同样的问题。例如,“运用大数据技术分析收集的证据”适用的核心问题是因为证据获取方式的不合法而导致证据不可采,而大数据证据同样会面临这一问题。因此,理清大数据证据的适用问题,对解决“运用大数据技术分析收集的证据”的可采性问题同样具有重要意义。

二、第一重困境:大数据证据与法定证据种类的不适应

(一)大数据证据的证据种类争议

我国《刑事诉讼法》第50条第2款明确规定:“证据包括:(一)物证;(二)书证;(三)证人证言;(四)被害人陈述;(五)犯罪嫌疑人、被告人供述和辩解;(六)鉴定意见;(七)勘验、检查、辨认、侦查实验等笔录;(八)视听资料、电子数据。”该条关于证据种类的规定实际上构成了我国证据审查与适用的第一道门槛,即只有符合法定证据种类要求的信息才有可能成为证据。对大数据证据而言,其在形式上的争议便成为首先要解决的问题。

(1)“证人证言说”。“证人证言说”是指将产生大数据分析报告的机器作为一种证人,进而将大数据证据列为证人证言。该观点认为,机器在解决法律争议事实中正扮演着越来越重要的角色,

机器传递出的一些信息可作为“机器证言”①。但这一观点主要存在两个问题:第一,其有可能剥夺当事人对质的权利。对质权是被指控人“讯问或业已讯问对他不利的证人,并使对他有利的证人在与对他不利的证人相同的条件下出庭和受讯问”②的权利。对质权可以有效地保护当事人的权利,同时帮助法官进行事实认定。但当证人变为机器时,被指控人则难以和冷冰冰的机器实现这种对抗。第二,机器证言的可信性将难以审查。根据“证言三角形”理论,证言的可信性需要从证人的感知能力、记忆能力、诚实性和叙述能力等方面进行综合分析加以判断[4]。但由于机器人不会有眼神、表情等状态变化,因此事实认定者很难对“机器证人”的这四项能力进行判别,自然也无法判别其证言的可信性。

(2)“鉴定意见说”③。“鉴定意见说”是指将大数据证据作为鉴定意见加以使用。该观点认为,从形式上看,大数据证据与鉴定意见均涉及科学问题,二者具有一定相似性。有学者指出:“最好把资金大数据分析纳入司法鉴定范畴。这有利于司法实践的展开,在法律上也可以找到依据。”[5]实践中,此观点也得到了一定程度的认可。例如某份判决书中曾提到,“安徽平泰司法鉴定所平泰司鉴字〔2018〕002 号司法鉴定意见证实:平台数据反映运营中心共发展会员 5 737 人,吸收会员投资 298 884 000 元,并造成其中 4 464 名会员损失 89 821 160 元”④。而另一份判决书中也提到了类似的观点:“重庆市科信电子数据司法鉴定所〔2017〕鉴字第 015 号《司法鉴定意见书》证明,李志超会员账号 WWac001、WWac002、WWac003 的下线层级、会员及获利的情况。”⑤这两个案件均是将大数据报告作为鉴定意见加以使用。相较于“证人证言说”,“鉴定意见说”更具合理性,这是因为大数据证据与科学证据本质上都是机器分析的结果。

然而,这一路径的困境在于,在传统鉴定意见中,机器主要承担的是工具性作用,鉴定意见的得出是鉴定人员根据机器的结果进行分析而实现的;而大数据证据则有所不同,大数据证据的获取并不需要专家参与,而是由机器运算产生。因此有学者认为,大数据证据“在很大程度上是由机器算法给出实质判断——不同于以往专家借助仪器设备作出判断,这对于以由专家作出判断的司法鉴定体制是一个过于超前的突破”[6]。

(3)“独立证据说”。“独立证据说”是指将大数据证据列为一种新的证据种类并通过法律加以确定。该观点认为,由于大数据证据具有不同于其他证据形式和种类的特点,因此在未来可以作为单独的证据种类。该观点对大数据证据在司法实践中的运用具有建设性意义。理由在于,目前的证据形式尚无法融洽地将大数据证据涵括其中。而大数据证据作为一种可能的证据形式,已经对目前审判工作产生了重要影响,单列其为一种证据形式似乎并无不妥。但该观点的问题是,立法活动较长的周期无法满足实践中对于大数据证据证据种类问题的迫切需求。众所周知,立法相对于司法活动而言具有滞后性,且程序较为复杂。只有在司法实践的需求达至一定程度后,立法活动才会展开。而当下,大数据证据已经在司法实践中开始运用,且大部分法官都是因大数据证据无法归类而对其持否定的态度。例如,笔者以“大数据分析报告”为题进行检索,查询到相关案例 78 件,而

①See Andrea Roth, Machine Testimony, 126 Yale L. J. 1972 (2017).
②《公民权利和政治权利国际公约》第 14 条第 3 款(戊)项。
③“专家辅助人说”与该观点类似,因此笔者将与之相似的观点皆列举在这一观点之下。
④参见安徽省高级人民法院刑事判决书(2019)皖刑终 118 号。
⑤参见安徽省灵璧县人民法院刑事判决书(2018)皖 1323 刑初 41 号。

在剔除同案多个当事人以及与大数据证据无关的案件后,所剩案件为34件,其中认定大数据证据的案件仅3件,大数据证据认定率不足10%⑥。这显示了大数据证据在司法实践中无法被认可的困境。如果立法不能尽快对该类信息达成共识,那么大数据证据的合法性仍将不断遭受质疑。

此外,不同的理论和实践对大数据材料的性质也给出了其他的观点。例如,有学者认为大数据分析报告是运用数据查询、比对和挖掘技术对收集到的大数据进行处理得出的关于案件事实的结论[7]。这实际上是将大数据报告视为一种书证。但大数据证据与书证之间存在着较大的差距:书证是案件发生过程中所形成的证据,而大数据证据是对案件发生过程中产生的数据进行大数据分析处理的结果,因此把大数据证据作为书证将引发大数据证据审查宽松化的风险。实践中,亦有判决将大数据材料作为侦查材料运用,如"根据通话详单、活动轨迹、大数据分析等手段发现'土匪'系邵阳县塘渡口镇江边村的龚某良,还经李某辨认予以确认,从而对其进行重点侦查,最终将其抓获"⑦。但按照这一方式处理,大数据材料将只能作为线索而不具有证据地位,这与鼓励采纳证据的证据法基本精神相背离。其实,该大数据分析报告完全可以作为认定"土匪"就是被告人的证据。

总之,大数据证据在证据种类问题上所面临的困境是:一种新的证据形式无法与传统证据形式相契合因而导致其无法获得合法的地位。

(二)大数据证据证据种类问题的出路:"三步走"策略

对于大数据证据的证据种类问题,应采取短期、中期、长期的"三步走"策略,有效规范大数据证据在法庭上的使用。

现阶段应将大数据证据作为鉴定意见。理由有二:一方面,大数据证据并非完全意义上的机器活动,其归根结底还是人类设计的算法运行的结果。只是相较于其他鉴定意见而言,人类参与的因素相对较少,但其本质却并没有太大差异。在实现一定程度的算法公开后,大数据技术并非不可运用科学技术加以鉴定,尽管整体的难度较大,但足以作为权宜之计。另一方面,大数据证据在法庭上的运用,必然需要专家证人加以配合。因此,将大数据证据作为鉴定意见,可以在最大程度地降低大数据证据的应用风险的同时,更容易为事实认定者所接受。

第二阶段应通过修法将大数据证据作为独立的证据种类。在当下算法公开等大数据技术尚且存在问题的前提下,长时间将其作为鉴定意见,容易引发司法实践中将大数据证据和一般科学证据相等同的错误定位。而基于前文的分析可以发现,大数据证据应该适用独立的标准。考虑到实践的可操作性以及可接受性,将大数据证据作为单独的证据种类更适宜。

第三阶段应逐渐放弃将证据种类作为证据门槛的做法。不可否认的是,依靠证据种类作为证据采纳的第一道门槛,在我国司法发展水平不足的前期对于规范事实认定工作具有一定作用。然而,从证据法的最基本的精神出发,任何具有相关性且符合法律要求的信息均可作为证据使用。毕竟证据的获取成本相对较高,轻易排除证据将直接影响事实认定的准确性。而规定证据种类的方法,颇有"法定证据主义"的嫌疑,其可能导致事实认定过程的形式化,以及程序正当性的虚无化[8]。本质上讲,只要大数据证据可以让事实认定者认为待证事实更可能或更不可能,且该证据没有应排

⑥在78份判决书中,海南省海口市秀英区人民法院民事判决书(2019)琼0105民初799号等44份判决书为当事人不同而判决内容完全相同的情况;北京知识产权法院民事判决书(2018)京73民终1134号为所提及的大数据与证据无关的情况。数据来源于中国裁判文书网,http://wenshu.court.gov.cn/,访问时间:2022年3月5日。

⑦参见湖南省邵阳市中级人民法院刑事裁定书(2018)湘05刑终25号。

除的其他情形,就应当被允许以证据的形式进入法庭之中。

三、第二重困境:因可靠性质疑而导致的相关性困境

(一)相关性困境的本质:证据的可靠性

传统相关性的认识是一个由证据性事实到推断性事实再到要素性事实的过程,而每一步的推断都需要概括进行连接,概括的组成主要是逻辑和一般经验,其基础是人类的社会"知识库",即从科学知识到流言蜚语[9]。但大数据技术却可以发现一些人类基于一般经验所无法发现的关联。例如,某地区在侦查贪污案件时,通过比对"去世人员"和"农村低保"两个数据库,有效地发现了原本未曾发现的有关贪污案件的疑点,而这些疑点是人类一般经验所难以发现的[10]。由此产生了此种疑问:这类人类通常不能通过一般经验解释的信息是否可以作为证据使用?特别是当大数据技术发挥其预测功能时,这样的矛盾和困境将更加尖锐。例如,《公安机关办理刑事案件程序规定》第224条规定:"执行拘留、逮捕的时候,遇有下列紧急情况之一的,不用搜查证也可以进行搜查:(一)可能随身携带凶器的;(二)可能隐藏爆炸、剧毒等危险物品的;(三)可能隐匿、毁弃、转移犯罪证据的;(四)可能隐匿其他犯罪嫌疑人的;(五)其他突然发生的紧急情况。"假设侦查人员运用大数据技术发现穿红色衣服的人更可能存在上述条文中的状况,并在没有搜查证的前提下对某一地区所有穿红色衣服的人进行搜查,那么这一大数据分析报告是否可以作为证据证明其行为的合法性?

这一例证实际上揭示了人与机器的差异,若按照传统的方法进行相关性分析,则很难单纯依靠一般经验发现衣服的颜色与携带凶器等紧急情况之间的联系。因此,有学者认为,"大数据分析方法让我们看到了瞬间大批量处理非结构化信息的可能性,同时大数据分析方法能够弥补人类对庞大数据分析理解上的不足,为事实认定者提供了基于数据的'数据经验'或者'特殊经验'"[11]。该说法在一定程度上揭示了大数据证据的相关性困境,但这种认识仍不全面。以红色衣服和携带凶器这一假设为例。其中,证据性事实是"大数据报告显示,穿着红色衣服的人更容易携带凶器";而要素性事实是"在没有搜查证的前提下,可以对可能携带凶器的人进行搜索",那么连接这两个事实的概括并非是"通常,穿红色衣服的人容易携带凶器",而是"通常,大数据分析报告是可靠的"。而如果以前者作为概括的话,那么得出的要素性事实或是推断性事实只能是"在衣服颜色与携带凶器的关系上,人类的认知与机器是相似的",而并不能得出真正的要素性事实。因此,对于大数据证据而言,如果法官可以做出"通常,大数据分析报告是可靠的"的概括的话,那么这一证据将影响其认知,进而产生某一项事实更加可能或更加不可能的判断——则该证据具有相关性。反之,若其做出的概括是"通常,大数据分析报告是不可靠的",那么这一证据便很难具有相关性。而所谓的"数据经验"或"特殊经验"也仅仅是探寻可靠性的一部分内容,即如果人类能够更好地解释"数据经验"或"特殊经验",便可以更好地判断大数据证据的可靠性。当然,除了可靠性之外,大数据证据的相关性同样要满足一般证据相关性的标准。

(二)大数据证据可靠性的困境

1. 大数据的黑箱化运行

当今,算法的不公开是原则,公开才是例外[12]。正是这种不公开的黑箱化运行,使得大数据技术本身的可靠性容易受到各方质疑。

(1)大数据技术是一种全新的技术,对于大多数人而言,其对大数据技术的算法构成不甚熟悉。

而出于对陌生事物的警惕性,人类更偏向于采取较为保守的态度。因此,诸如算法本身是否科学,算法的结果是否准确等疑问便会不断产生。从实践中看,这种担忧有较大的合理性。因为算法由人类设计,其本身不可能做到绝对的客观中立,必然会受到一些因素的影响。而实践中所发现的这类问题,将进一步加剧人类因对于算法的不了解而导致的不安和质疑。

(2)算法的黑箱将导致参与性的缺失。由于人们无法参与整个决策形成和制定的过程之中,因此无法对决策提出自己的意见和建议。而这种缺乏参与和商讨的过程很有可能引发公信力危机。以城市公共空间治理为例,研究表明,公民参与城市公共空间治理有利于构建以人民为中心的城市公共空间,进而重建城市公共空间的公共性。反之,城市空间的公共性便会受到破坏[13]。再如,我国司法改革过程中不断通过人民陪审团等方式加强公众参与,也正是希望增强公众对司法的信任,从而提升司法认同。而当公众无法参与其中时,其自然容易对运行的结果产生怀疑。同理,算法的黑箱亦是如此。

(3)算法的黑箱化运行可能因信息不对等而引发怀疑。法律决策本身具有"透明化"的要求[14],其中一个重要的原因便是为了避免信息的不对等而导致权力的异化。申言之,由于信息不对等,信息优势的一方可以利用各种方式来引导信息劣势的一方作出错误的判断。例如,有研究表明,信息不对称可能导致媒体为了寻求直接"非生产性"的经济租金(即媒体寻租),向社会作出不实或欺骗性报道,使优质企业在民众心中的诚信指数下降,信用和声誉受损[15]。因此,出于对巨大信息差的畏惧,黑箱化的运行容易受到各方的质疑而导致失信。以政府公信力为例,研究表明,作为政府信息的劣势方,不充分的信息公开可能使公众面临着在政治上的"逆向选择",他们将会对政府报以冷漠、不信任的态度[16]。同样,大数据技术如果不能有效公开,也会面临无法被信任的危险。

(4)算法黑箱化运行对诉讼程序的挑战可能进一步引发可靠性的质疑。在质证阶段,被指控人有可能因不了解大数据技术的运行原理而导致质证权利的不彰。质证权是被指控人一项重要的权利。从程序意义上讲,质证有利于增强当事人的诉讼参与性;从实体意义上讲,质证可以让事实认定者更好地了解证据,进而提升事实认定的准确性。质证的主要内容是对证据的证明力、可信性(可靠性)等内容进行质疑,进而阻断事实认定者的经验推论链条。而大数据的黑箱化运行使得被指控人根本无从了解该证据的信息,便也无法对其证明力、可信性(可靠性)等内容作出判断。在认证阶段,大数据证据可能会引发法官事实认定的困境。根据"证据之镜"原理,对事实的认定只能借助证据这一桥梁实现[9]。但大数据的黑箱化运行导致法官无法对大数据证据有最基本的了解,并且黑箱化的方式也限制了鉴定人或者专家辅助人对大数据证据进行解释的空间。这就导致了法官无法对大数据证据的可靠性进行判断。

2. 大数据技术的复杂性

影响大数据证据可靠性的另一个重要因素是大数据技术的复杂性。大数据技术打破了原有的时空界限,带来了前所未有的突破。与此同时,大数据技术也引发了个体行为向集体行为的靠拢。这主要是由于大数据技术十分复杂,单靠个人的力量难以实现[17]。而这种复杂性对于从事法律活动的"外行人"而言更是难上加难。具体而言,"外行人"理解大数据技术的困境主要在于:(1)内容的复杂性。大数据技术所涉及的算法框架等内容对于很多外行而言都是陌生的。例如,在法学领域,对于大数据的概念等有关内容有很多无法达成一致,而造成这种不一致的原因与论者对技术的不了解有较大关联。(2)计算过程的复杂性。对于整个大数据技术而言,一般人的了解只能局限于

两个端口:开端——对于大数据技术的特点有一定认识;末端——知晓大数据运算后的结果。以大数据分析报告为例,大数据技术基于何种方式生成的分析报告,无论是使用者还是事实认定者都很难清晰而准确地认识。(3)结果的理解困难。大数据材料结果理解的困境主要在于一些结论难以基于一般经验解释。例如,大数据技术发现尿布与啤酒的销量有一定关系,对于这一结果的认知难度并不在于读数——了解尿布与啤酒有一定关联;而在于理解和接受这种关联——尿布和啤酒为何会有关联⑧。

(三)大数据证据可靠性困境的出路

大数据证据是否具有相关性的关键在于其是否具有可靠性。如前文所述,大数据证据可靠性的主要困境在于算法的"不透明性+复杂性"。因此,解决大数据证据可靠性困境的关键在于如何在保证准确性的前提下,最为简单地公开算法。从完整性的角度讲,公布整个大数据运行的全过程显然对于缓解黑箱问题最为有力。然而,尽管这种方式对于法学家而言十分理想,但对于技术的拥有者而言则并非如此,有些算法十分容易导致行业内效仿而造成利益的损失。法律的制定需要权衡各方的利益,基于不同价值考量进行权衡分析。例如,尽管准确的事实认定是证据法的重要追求,但为了维护社会的和谐稳定,依然出现了不得用以证明过错或责任的规则。同样地,在信息公开需求最强烈的政府信息公开问题上,《中华人民共和国政府信息公开条例》依然设立了例外款项。因此,大数据证据的公开同样应该设立一定限度,以保障各方的利益。

由此便引发了下一个问题,即该限度的标准为何。这要首先考量事实认定活动所面临的主体问题。就事实认定而言,事实认定者仅仅需要的是理性的人,并不需要具有任何专业知识。理由在于,事实认定的过程主要依赖于人类的逻辑和一般经验。即使是法官,其社会"知识库"也并不必然优于一般人。因此,算法公开的限度应以事实认定者即理性的普通人为考虑对象。基于此,笔者认为判断大数据证据公开的合理方式应该是公布其所依据算法的历史准确率,理由如下:(1)该方式提供了一个清晰的标准——数字,让理性的普通人足以进行判断。实际上,通过数字解读证明问题的方式早已有所应用,例如,将"排除合理怀疑"的标准与95%这一数字画上等号后,便可以帮助事实认定者更好地明确"排除合理怀疑"这一较为模糊的证明标准。同样,在大数据证据公开问题上运用数字化的标准可以起到清晰化的作用。(2)该方式提供了一种较为容易理解的标准。对于事实认定者而言,其只需要对数字的大小进行判断。这实际上是将较为复杂的对于算法可靠性的判断转变为较为简单的对于数字的判断。考虑到大数据技术的复杂性,除了专业人士之外,算法的公开对于一般的事实认定者而言没有任何意义[18],甚至可能因分散其注意力导致事实认定走向歧途。(3)该方式提供了一种较为有效率的标准。为了避免让当事人的权利长期处于不安定的状态,诉讼活动必须在一定时间内进行,这也是法律对于诉讼活动期限规定的意义。同样地,大数据证据可靠性的判断方式如果过于复杂,将导致法庭运用大量的时间在此问题之上,从而造成不当的拖延。而对历史准确率的分析和判断则相对容易,并不需要事实认定者用大量的时间去理解和学习其中复杂的过程,显然有利于效率价值的实现。

需要注意的是,算法历史准确率的真实性必须通过制度进行背书。换言之,应当通过相应的规

⑧此为大数据技术应用的经典案例,沃尔玛运用大数据技术对消费者购物行为进行分析后发现其具有关联,便将二者摆在一起进行售卖,而这一举措使二者的销量均大大增加。

定与监管来保证算法历史准确率的可靠性。具体而言,首先,算法历史准确率公布的主体应该是算法开发者(或改进者),因为开发大数据算法的一个组成部分便是计算(改进)正在进行的算法的准确性。其次,应当由政府部门牵头,依托具有相应专业人才、技术的支撑和监管能力的行业自律组织负责算法的监管[19]。这样既可以有效保证算法监管的权威性,又可以缓解政府部门人力不足的困境。监管部门应当对算法的准确性制定相应的评价标准,进而监督算法历史准确率的真实性。最后,应当建立算法备案制度。算法开发者应及时向算法监管部门就算法的基本情况进行备案,并且定期对算法的历史准确率等内容进行上报。此外,监管部门也可以运用区块链技术,对算法的有关信息进行存证和监管。

然而,如果仅仅依靠历史准确率进行判断,难免会给人产生这样的疑问:前100次的准确计算是否可以保证第101次计算的准确?这便需要鉴定人、专家辅助人进行解释,使一般人能够理解基于"数据经验"产生的关联,从而进一步对大数据证据的可靠性进行判断。

四、第三重困境:对隐私权的侵犯和"证据偏在"导致的可采性困境

(一)证据可采性困境的具体表现

1. 对隐私权的侵犯

可采性规则为了保证公民的基本权利,往往对侵犯公民基本权利的证据予以排除。例如《刑事诉讼法》第56条规定:"采用刑讯逼供等非法方法收集的犯罪嫌疑人、被告人供述和采用暴力、威胁等非法方法收集的证人证言、被害人陈述,应当予以排除。"这实际上就是出于保障公民基本的人身权利而对获取方式不符合程序要求的证据予以排除。

而大数据证据的获取可能直接对公民的隐私权造成威胁。以大数据侦查为例,近年来,侦查机关通过各种方式不断提升数据扩充的可能性,甚至以相关数据总量的多少进行计较。例如,"日均采集'一标三实'等重要基础信息30余万条"[20]式的表达更多被作为"标语"而为侦查部门所宣传。正是由于这种信息获取的不断扩张,致使公民隐私权的边界被不断缩减,算法系统的关联技术甚至"知道你想做什么","数据主宰世界"的隐患正在侵蚀用户信息生态环境[21]。换言之,无论是否有意为之,大数据证据的获取者在提取与案件有关的信息时,往往会获得与案件无关的信息进而侵犯公民的隐私权,而且这种对于隐私权的侵犯很多时候无可避免。在大数据海量数据的要求下,"全数据"的追求使得数据收集者在获取数据时为了保证准确性,大多会优先考虑如何尽可能地获取更多数据,而不去考虑该数据的获取是否合法或是否侵犯公民的隐私权。这势必带来一个十分严峻的问题——用于作为证据的大数据信息是否"干净"。

此外,数据获取后的挖掘将进一步加剧这种对隐私权的侵犯可能性。如前文所述,大数据技术往往可以利用一些看似不相关的信息挖掘出全新的知识,而这些新的知识对于隐私权的侵犯同样不容忽视。例如,美国在线AOL在2006年曾公布了3个月近2 000万条真实的搜索记录,搜索的内容很可能涉及个人隐私的敏感信息,与特定用户有着密切的联系。诸如"尿布"这样的搜索,可以让人轻易地推断出用户是一名婴儿的父母[22]。可见,大数据挖掘技术将进一步加剧对公民隐私权的侵犯。

正义是证据法的根本价值追求,如何保障个人的权利不被侵犯便成为证据规则所考量的重要内容。为了这一目的,准确的价值往往要让位于正义的价值。而大数据证据的获取的一方十分有

可能侵犯公民的基本权利——隐私权。出于对这种行为的警惕,大数据证据将有可能因不满足可采性规则而被排除在法庭之外。

2.“证据偏在”的存在

“平等武装”和“平等对抗”是现代刑事诉讼的基本追求。自欧洲人权法院第一次通过判例将该原则予以明确后,各主要国际刑事法院(法庭)均将其作为重要的程序性原则[23]。然而,由于诉讼的双方可能在人力物力等方面不平等,诉讼过程很难达至完全意义上的平等。一个不容忽视的事实是,个人获取证据的能力和方法相较于国家或是大企业显得远远不足,由此可能产生“证据偏在”的现象,即因诉讼双方获取证据能力上的差异导致所获取的证据更多被一方持有。而这一现象将有可能直接影响公民的权利。例如,在萧山冤案中,警方并未将有利于被告人的证据同其他证据一并移交至公诉机关,直接导致本案成为冤假错案[24]。“证据偏在”现象的存在,十分容易因一方隐瞒部分证据而导致事实认定者产生不公正偏见、误解等危险。从可采性规则的角度出发,当可能引发的不公正偏见等危险性实质上超过证明力时,该证据便不再具有可采性。

大数据证据的获取无疑进一步强化了这种“证据偏在”现象。如前文所述,大数据技术是一种十分复杂的,需要多人协作才能实现的技术。而大部分诉讼当事人都不具有这样的能力来支持完成这项工作,故大数据证据只能掌握在少数人手中。以李某、叶某安、徐某奎等集资诈骗案为例,本案中,平台数据反映运营中心共发展会员 5 737 人,吸收会员投资 298 884 000 元,并造成其中 4 464 名会员损失 89 821 160 元⑨。拥有这样计算能力的主体往往是公权力机关或是大型企业,这使得大数据证据的获取实际上面临着三重危险:(1)隐藏或篡改证据的风险;(2)数据独家解释的风险;(3)无法质证的风险。这三重风险不仅可能导致事实认定者产生错误的偏见,更可能导致因专业性和排他性过强而使得当事人无法行使质证权等基本诉讼权利。结合大数据证据本身可靠性的质疑,其完全有可能因危险性过高而被排除。

(二)大数据证据可采性困境的出路

1. 构建“原则+制度+技术”的融合规制路径

当下,大数据收集活动中的操作流程较为混乱,为保障公民的权利不受侵犯,应当构建“原则+制度+技术”的融合规制路径,保障大数据证据的合理获取。

从原则角度出发,大数据技术的应用应遵循以下原则:(1)数据有限使用原则。在运用大数据技术时,其收益应高于对权利的损害。这既要求数据使用者在使用数据时充分评估大数据技术的利弊;也要求其在运用大数据技术时尽可能保持克制。如果有替代性信息,则不应侵犯公民的隐私信息;如果必须使用公民的隐私信息,也应最小限度地运用,并尽可能通过技术手段避免信息的泄露。(2)数据主体“弱同意”原则。在数据主体的“强同意”模式下,数据主体具有充分的信息自决权。但该模式并不利于数据的流通,而且大数据技术立足于海量数据分析,要求数据使用者与每个人进行谈判也不具有可操作性。而“弱同意”原则则构建了一种“情境合理+拟制同意=合法处理”的数据使用模式。该模式在保护公民隐私权的同时,也维护了大数据技术的应用价值[25]。(3)数据甄别原则。放弃目前对大数据不加甄别全面搜集的状态,转而只搜集与案件有关的信息。例如,在搜集某人在某网站所发布的信息时,不应不加甄别地将网站内所有用户发布的内容和所有用户

⑨参见安徽省高级人民法院刑事判决书(2019)皖刑终 118 号。

的信息都加以使用,而是仅针对要调查的个人或事件进行搜集,尽可能保障不侵犯其他人的权利。

从制度角度出发,完善大数据技术的监管体制。第一,构建大数据技术风险评估系统,对大数据技术的应用进行风险等级评定,严格限制高风险大数据技术的引用,并通过程序手段对风险加以控制[26]。第二,引入大数据技术应用的审查机制,其主要由两条路径构成。第一条路径是大数据监管机构的审查。如前文所述,大数据技术的使用应向监管机构备案,而监管机构可以就大数据技术的使用情况进行事后审查;第二条路径是司法审查。从长远来看,伴随着各类强制性措施司法审查机制的陆续建立,由更为中立的法官统一对情报信息领域与刑事司法领域行使事先审批权则是更为合理的选择[27]。

从技术角度出发,尝试通过更加先进的技术完善隐私保护机制。例如,"数据脱敏"技术便对隐私权的保护具有重要意义。该技术可以在保留原始数据的业务价值、技术价值的前提下,对敏感信息进行脱敏、隐蔽处理[28]。常见的脱敏手段包括替换、截取、加密、掩码等。当然,仅仅依靠"数据脱敏"等技术并不足以完全保护公民的隐私权。例如,脱敏后的隐私数据可以通过技术实现再复原。对此,一方面应继续提高技术水平;另一方面同样应把这类技术纳入法律的监管之下。

2. 增强诉讼的对抗性

在涉及大数据证据的案件中,诉讼双方很难处于一种均势的状态,特别是在刑事诉讼中,这种对抗性的缺失更加明显。为了维护诉讼中"平等对抗"的原则,有效保护被告人的诉讼权利十分必要。这种保障可以通过两种方式实现:第一,完善证据开示制度。证据开示是当事人主动向对方寻找证据和信息的一种权利,即要求对方当事人出示信息的诉讼行为[29]。由于在大数据证据出现的案件中,双方当事人往往很难有效抗衡,那么最好的方式便是让双方共享彼此所持有的证据信息。而如果能够保证所有的证据信息都被分享,那么"证据偏在"的问题在某种程度上就会得到解决。因此,在有关大数据证据的诉讼中,双方当事人应当在证据交换环节公开彼此的证据。这同样涉及大数据证据公开限度的问题,如果说前文所述的算法历史准确率的公开是为了保障大数据证据的可靠性,那么为了避免大数据运算过程中对公民基本权利的侵犯,保障大数据证据的合法性,大数据证据的提出者需要同时说明所运用的算法中包含的影响因素。当这些影响因素涉及权利侵犯、歧视等不可接受的内容时,这一算法同样不可接受[39]。第二,提升对于大数据证据运用的监督。通过有效的监督可以在一定程度均衡双方当事人的力量。这种监督主要来自两个主体:第一个主体是检察机关作为法律监督机关的监督。为此,检察机关应建立有关的证据数据库,这一数据库主要有两方面价值,一方面,这一数据库可以用于监督有关大数据证据获取过程中存在的问题,通过有效的数据分析和挖掘,判断各项大数据证据的合法性;另一方面,在刑事诉讼中,这一数据库可以挖掘当事人无罪的证据。这种方式可以将"权力—权利"的对抗有效地转变为"权力—权力"的对抗。更为重要的是,随着数据量的不断增加,在发现当事人有罪的可能性增加的同时,发现其无罪的可能性也同样在增加。这无疑为冤假错案的平反提供了一个有力的方式。这一方式也与西方所提"数字无罪"的概念有很大的相似性[31]。第二个主体是社会主体的补充监督。不容忽视的是,检察机关在刑事诉讼中承担着公诉人的任务,因此其可能无法完全摆脱自我监督的困境。这就需要社会第三方主体的补充监督。社会主体之所以可以发挥监督作用,主要是由于大数据权力重塑了传统权力的结构功能、组织形态和运行机理,催生了大数据驱动式社会监督模式[32]。随着社会各个主体数据获取能力的增强,其完全可以起到补充监督的作用,即通过社会主体自身的数据库,对诉讼

活动中大数据证据获取的合法性进行检查,同时为弱势一方提供相应的证据。随着技术的发展,信息的隐瞒与控制将越发困难,这也给社会主体进行监督提供了客观的环境。

大数据证据对于提升事实认定的准确性具有重要价值,但其也对传统证据规则造成了一定程度的冲击。为了更好地保证大数据证据在法庭中正当合法的应用,适度地完善相关规则十分必要。但需要注意的是,当下的实践更多看到的是大数据技术的价值,而忽视了其可能带来的弊端。因此,大量的实践操作都无法通过程序性的检验,也同样无法作为证据在法庭中加以应用。从某种意义上讲,规范取证行为是大数据材料向大数据证据转化的关键。

参考文献:

[1]丰叶. 职务犯罪大数据证据研究[J]. 科技与法律,2020(1):76-85.

[2]谢君泽. 论大数据证明[J]. 中国刑事法杂志,2020(2):125-137.

[3]王燃. 大数据时代侦查模式的变革及其法律问题研究[J]. 法制与社会发展,2018(5):110-129.

[4]张保生. 证言三角形及其理论意义[J]. 中国政法大学学报,2015(2):2,161.

[5]何家弘,邓昌智,张桂勇,等. 大数据侦查给证据法带来的挑战[J]. 人民检察,2018(1):54-57.

[6]刘品新. 论大数据证据[J]. 环球法律评论,2019(1):21-34.

[7]胡铭,龚中航. 大数据侦查的基本定位与法律规制[J]. 浙江社会科学,2019(12):12-20,55.

[8]孙远. 论法定证据种类概念之无价值[J]. 当代法学,2014(2):99-106.

[9]张保生. 事实、证据与事实认定[J]. 中国社会科学,2017(8):110-130,206.

[10]庞岚. 官员有 11 套房自以为安全,不料这举动被大数据揪出[EB/OL]. (2018-05-29)[2020-3-12]. https://finance. sina. com. cn/china/gncj/2018-05-30/doc-ihcffhsv2868843. shtml.

[11]周蔚. 大数据在事实认定中作用机制分析[J]. 中国政法大学学报,2015(6):64-82,160.

[12]徐凤. 人工智能算法黑箱的法律规制:以智能投顾为例展开[J]. 东方法学,2019(6):78-86.

[13]陈水生,屈梦蝶. 公民参与城市公共空间治理的价值及其实现路径:来自日本的经验与启示[J]. 中国行政管理,2020(1):135-141.

[14]左卫民. 关于法律人工智能在中国运用前景的若干思考[J]. 清华法学,2018(2):108-124.

[15]李杰. 论非对称信息下媒体寻租对信用与经济的影响[J]. 首都师范大学学报(社会科学版),2019(6):53-62.

[16]赵超,贺华. 信息不对称理论下政府公信力影响机理探析[J]. 西北工业大学学报(社会科学版),2010(4):6-10,34.

[17]贾向桐. 大数据的新经验主义进路及其问题[J]. 江西社会科学,2017(12):5-11.

[18]张淑玲. 破解黑箱:智媒时代的算法权力规制与透明实现机制[J]. 中国出版,2018(7):49-53.

[19]孙清白. 人工智能算法的"公共性"应用风险及其二元规制[J]. 行政法学研究,2020(4):58-66.

[20]何仕杨. 最强实战|四川公安大数据规划建设见成效[EB/OL]. (2019-07-05)[2020-3-12]. http://sc. china. com. cn/2019/jinri_0705/327588. html.

[21]纪楠,李平. 算法时代用户隐私权的保护[J]. 青年记者,2019(26):78-79.

[22]孙广中,魏燊,谢幸. 大数据时代中的去匿名化技术及应用[J]. 信息通信技术,2013(6):52-57.

[23]王秀梅,陈朗. 论国际刑事辩护"平等武装"原则[J]. 刑法论丛,2014(2):389-406.

[24]史炜,王逍奕. 侦控机关刑事案卷移送中的"证据偏在"[J]. 广西警察学院学报,2019(4):43-47.

[25]蔡星月. 数据主体的"弱同意"及其规范结构[J]. 比较法研究,2019(4):71-86.

[26]张衡. 大数据监控社会中的隐私权保护研究[J]. 图书与情报,2018(1):71-80.

[27]程雷. 刑事司法中的公民个人信息保护[J]. 中国人民大学学报,2019(1):104-113.

[28]王毛路,华跃. 数据脱敏在政府数据治理及开放服务中的应用[J]. 电子政务,2019(5):94-103.

[29]白绿铉. 美国民事诉讼法[M]. 2 版. 北京:经济日报出版社,1998:78.

[30]SIMMONS R. Quantifying criminal procedure: How to unlock the potential of big data in our criminal justice system[J]. Mich. St. L. Rev. ,2016,4:947-1018.

[31]FAIRFIELD J,LUNA E. Digital innocence[J]. Cornell Law Review,2014,99:981-1076.
[32]蔡玉卿. 大数据驱动式社会监督:内涵、机制与路径[J]. 河南社会科学,2019(8):52-58.

Triple dilemma and solutions for the application of big data evidence

ZHENG Fei[1], MA Guoyang[2]
(*1. Law School, Beijing Jiaotong University, Beijing 100044, P. R. China;*
2. Collaborative Innovation Center of Judicial Civilization, China University of Political Science and Law, Beijing 100088, P. R. China)

Abstract: Big data evidence is the evidence used in the trial to screen, summarize, refine, and form a conclusion on the massive data. Big data evidence is different from evidence analyzed or collected by big data technology. The latter does not pose a significant challenge to the traditional evidence rules, while the former leads to the maladjustment between big data evidence and traditional evidence rules, which leads to the triple dilemma of using big data evidence in court. The first dilemma is the inadaptability between the types of big data evidence and the types of legal evidence, which should be solved gradually through the three-step strategy in different periods. In the first stage, big data evidence should be regarded as an expert opinion. In the second stage, big data evidence should be regarded as an independent type of evidence. In the third stage, the practice of taking the type of evidence as the threshold of evidence review should be abandoned. The second dilemma is the relevance dilemma caused by reliability, which is due to the black box operation of big data and the complexity of big data technology. The simplest and direct solution is to disclose the historical accuracy of the algorithm. Among them, the main body of publishing the historical accuracy of the algorithm should be the algorithm developer (or improver), because an integral part of developing big data algorithm is to calculate (improve) the accuracy of the algorithm in progress. At the same time, in order to ensure the credibility of the historical accuracy published by algorithm developers (or improvers), government departments should also take the lead and rely on industry self-discipline organizations with corresponding professional talents, technical support and supervision ability to supervise the algorithm. In addition, if necessary, appraisers and expert assistants should be sought to explain, so that ordinary people can understand the relevance based on "data experience", so as to further judge the reliability of big data evidence. The third dilemma is the admissibility dilemma caused by the invasion of privacy and the influence of "evidence bias". This dilemma should be solved by constructing the integrated regulation path of "principle + system + technology". From the perspective of principle, the application principles of big data evidence include the principle of limited use of data, the principle of "weak consent" of data subjects and the principle of data screening. From the perspective of system, on the one hand, a big data technology risk assessment system should be built to assess the risk level of the application of big data technology. On the other hand, the review mechanism for the application of big data technology should be introduced, including the review of big data regulators and the review of judicial organs. From a technical point of view, the privacy protection mechanism through more advanced technologies such as "data desensitization" should be tried. In addition, the resolution of the third dilemma of big data evidence also needs the enhancement of adversary of litigation by improving the evidence discovery system and other methods.

Key words: big data; big data evidence; evidence application; fact finding; triple dilemma

(责任编辑 胡志平)