算法侵害责任框架刍议

王　莹*

内容提要　算法决策的主体多元性、开放性、不透明性、自主性等技术特征给算法侵害归责带来挑战，有必要在检验与拓展传统侵权法与刑法责任框架基础上，结合算法技术特征构建一个层级式的算法责任框架。对于产生具体侵害结果的算法损害归责而言，分别设置因果支配型的责任与非因果支配型的义务型责任。前者是指针对具有理解与控制技术可能性的算法适用过错责任与严格责任，后者是指针对不具有理解与控制技术可能性的算法适用义务型责任，为算法设计者与应用者设置避免算法侵害风险现实化的义务或对用户、社会公众法益的保护义务及相应责任。对算法妨害而言，因无具体权利侵害结果因而不存在结果归责问题，仅需对该算法妨害行为加以禁止，设置相应行政违法行为，或设置类似侵权法的公共妨害甚或刑法上的抽象危险犯进行追责。

关键词　算法侵害　算法妨害　算法损害　义务型责任

目　次

*　中国人民大学法学院未来法治研究院、刑事法律科学研究中心副教授，法学博士。本文系2018年度国家社科基金一般项目“信息刑法时代信息犯罪归责与治理模式研究”（项目批准号：18BFX101）的阶段性成果。

算法在带来效率、便捷的同时，其蕴含的技术风险也可能给个人与社会带来广泛的负面影响甚至侵害。一方面，算法决策可能因数据偏见、设计缺陷等原因导致歧视性或错误的决策结论，损害决策者或决策对象的利益，造成算法侵害。另一方面，算法技术的高度专业性、不透明性及决策数据的庞杂性等妨碍决策者及决策对象理解算法决策机制，大大增加了寻找损害原因及追究责任的难度。目前国内外研究主要集中于算法决策的事前规制领域，即数据赋权的源头规制方式与引入算法程序性审查与评估等方式，而对算法的事后规制，即对算法造成的侵害进行归责、追责并通过责任制度进行算法侵害的防治，研究则较为薄弱。

随着算法侵害的风险日益显现，欧盟与美国开始重视算法与 AI 领域的责任问题。2020 年 10 月，欧洲议会通过《欧洲议会关于欧盟委员会就人工智能民事责任制度建议的决议》（以下简称《AI 民事责任决议》）①，呼吁以法规的形式为人工智能（AI）系统确立新的普遍规则，强调 AI 系统导致侵害的责任问题是构建 AI 法规框架的关键一环。我国在关注算法事前规制的同时，也亟需加强对算法事后规制的研究。

一、算法侵害分型：算法妨害与算法损害

近年来，算法的本土应用不断带来挑战与风险。例如，河北邯郸特斯拉自动驾驶首例交通事故引发对算法错误的思考、② 携程大数据杀熟反映算法歧视质疑，③ 美团外卖骑手案中的绩效算法则折射了运用算法作出不利于被决策对象的决策、对其进行干预的算法操纵隐忧④。为了防范算法歧视、大数据杀熟、诱导沉迷等算法不合理应用对传播秩序、市场秩序和社会秩序的扰乱，应对其给意识形态安全、社会公平公正和网民合法权益带来的挑战，2022 年 1 月 4 日国家互联网信息办公室、工业和信息化部、公安部和国家市场监督管理总局四部门联合发布《互联网信息服务算法推荐管理规定》（以下简称《算法规定》），规范算法推荐服务，规制广泛、多维的算法侵害。

算法侵害不同于传统物理社会的权利或法益侵害：首先，上述侵害有些无法落入传统部门法既定的权益框架，因而无法根据传统部门法责任框架追责。其次，基于算法决策机制的不透明性、专业性、复杂性，算法侵害的产生过程更为复杂、隐秘，人的因素与机器的因素叠加，给归责与追责带来挑战。因此，有必要超越传统部门法的界限，从法律与技术融合的视角对算法侵害进行分型，整合传统教义学责任理论进行针对性责任探讨。

---

① *See P9_TA (2020) 0276 European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on a Civil Liability Regime for Artificial Intelligence (2020/2014(INL))*, at https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_EN.pdf., Art.5. (Last visited on January 15, 2022)

② 参见苏宇：《算法规制的谱系》，载《中国法学》2020 年第 3 期，第 167 页。

③ 参见高富平、王苑：《大数据何以“杀熟”？——关于差异化定价法律规制的思考》，载《上海法治报》2018 年 5 月 16 日，第 B06 版。

④ 参见赖祐萱：《外卖骑手，困在系统里》，载《文摘报》，2020 年 9 月 12 日，第 1 版；丁晓东：《论算法的法律规制》，载《中国社会科学》2020 年第 12 期，第 139 页。

（一）算法妨害：基于风险或过程的抽象性侵害

美国学者 Balkin 对算法侵害进行了开创性的研究，将算法社会企业大规模适用大数据分类、风险评估技术给个人身份、声誉及行为带来的累积性消极影响类比为侵权法上的环境污染公共妨害，从而提出算法妨害概念（algorithmic nuisance）。算法妨害包括大数据风险评估对个人进行风险群体归类而带来的声誉侵害（如将个人贴上犯罪风险或财务风险的标签，笔者称之为算法标签）、对风险个体进行差别化对待而造成歧视（算法歧视）、压制或改变人们的数字行为轨迹而影响数字行动自由（算法归化）、利用算法操纵人们作出不利选择（算法操纵）、缺乏透明性及正当程序（算法黑箱）等。[⑤] 算法的声誉侵害、归化、操纵、程序瑕疵等皆是算法本身蕴含的技术风险或过程瑕疵，其造成的侵害或者是过程性的，或者是抽象的、模糊的，并未产生实际的损害结果，尚未直接侵害传统法律的某种既定法益或权利。其中，算法标签是指对个人或集体进行大数据画像，形成便于算法或网络平台推销不同产品或服务的电子身份标签，[⑥] 并存档、流通于数字空间，但并未构成名誉权侵害。算法归化是指广泛应用的算法限缩人们的行动自由空间，例如公共场所人脸识别算法、办公空间职员行为监控及绩效评估算法虽然并未达到侵犯隐私或限制行动自由的程度，但会在一定程度上干预人们的活动，被算法监控的人们会避免进入被监控的区域或在不得不进入时收敛或调整自己的行为。算法操纵是指算法通过对信息数据的生成合成、个性化推送、排序精选、检索过滤、调度决策等对个人或群体所接受的信息进行干预，并据此影响其意志形成与行为选择。例如，根据用户兴趣偏好过度推荐，造成“信息茧房”及诱导沉迷；利用算法注册虚假账号进行流量造假、刷量控评、饭圈互撕；通过屏蔽信息操纵热搜、榜单或检索结果排序等 [⑦]；配送算法设置严格的时间计算规则，对骑手进行剥削操纵。上述侵害从个案来看较轻微，但因算法的广泛应用具有集合性、累积性、系统性，可能对社会集体造成广泛而深远的负面影响。

（二）算法损害：基于结果的具体侵害

算法自动决策不仅导致过程性的、抽象性的算法妨害，也可能造成对传统部门法既定权益的直接侵害。算法的设计与应用是一个涉及众多技术主体与数据处理环节的复杂过程，算法可能因数据偏见、设计缺陷、编码错误或被第三人滥用而作出错误决策结论，从而导致具体的损害后果。[⑧] 算法也存在被第三人滥用的可能性，如微软开发的线上聊天机器人 TAI 能够通过学习对话内容进行自我迭代，一些推特用户向其发送种族主义

---

⑤ *See* Jack M. Balkin, *2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data*, Ohio State Law Journal, Vol.78:1217, p.1237–1239 (2017).

⑥ 如，大数据分析根据数据主体经济脆弱程度对其评分、归类、标签，并出售给金融公司，便于后者向其推销小额贷或高利贷等其他非传统金融产品。*See* Article 29 Working Party's Guidelines on Automated Individual Decision–making and Profiling for the Purposes of Regulation 2016 /679", at https://ec.europa.eu/newsroom/article29/item–detail.cfm?item_id=612053, p.10 (Last visited on January 15, 2022).

⑦ 参见《大数据杀熟如何解？诱导用户沉迷？专家聚焦算法新规五大看点》，载澎湃新闻网，https://www.thepaper.cn/newsDetail_forward_16167906，2022 年 1 月 15 日访问。

⑧ *See* Mark A. Lemley & Bryan Casey, *Remedies for Robots*, University of Chicago Law Review, Vol.86:1311, p.1323–1324(2019).

等有害信息，引导其发表纳粹言论。⑨

如果说算法妨害是算法技术过程的风险及消极影响，更侧重算法行为端，则算法歧视、算法错误、算法滥用类型的侵害则更侧重结果端，会引发具体的权益侵害，可称之为算法损害。算法损害的具体损害结果虽然可能表现为侵权法或刑法的权利或法益侵害，但基于机器学习算法的复杂性、不透明性、开放性与迭代性技术特征，损害的确定、因果追溯与归责更加复杂，给传统责任框架带来挑战。

上述抽象侵害与具体侵害的二分法将算法侵害分为算法系统技术过程蕴含的风险与该风险在具体场景下的实现即具体的权利侵害结果。如果将算法自动决策视为一个行为整体，可将其划分为行为过程及行为结果：抽象的算法妨害属于行为端，具体的算法损害属于结果端。这一区分为我们借鉴传统法教义学资源中的归责原理开展事后规制——算法侵害责任的研究提供了方法论路径。算法妨害因不存在具体的损害后果，因而不产生结果归责问题，可将算法妨害行为规定为侵权法或刑法中的行为构成要件并配置责任；算法损害则需要解决结果归责问题，即将结果归属于哪一个行为人或行为。

二、算法事前规制方案及其缺陷

针对大数据给个人带来的侵害隐私风险，Lessig 归纳了四种应对措施：法律、规范、市场与技术架构或代码。就法律措施而言，大致可分为事前和事后两种应对机制。⑩ 前者是预防性的，侧重于侵害发生前的防控；后者是回应性的，在侵害发生后对侵害进行归责、追责与救济。Lessig 代码理论的兴盛很大程度上催生了知情同意、通过设计的隐私保护、数据赋权等制度工具，但其法律规制中隐含的责任措施却并未在网络与人工智能法中获得应有的关注。针对算法自动决策的风险及侵害，目前学界主要从规制数据及自动决策过程两方面入手，提出事前的规制方案，但这些方案逐渐凸显不足。

（一）源头数据规制路径 / 个人赋权制衡的规制路径之不足

2018 年 5 月，欧盟出台《一般数据保护条例》，规定个人数据的一般原则、数据主体的权利、数据跨境转移、独立监管机构、权利救济及责任等。我国《个人信息保护法》也采取了从算法决策的数据源头管控与个人数据赋权制衡的策略，弥补了传统部门法规制的不足，为应对算法风险及侵害提供了重要法律制度保障。

1. 数据源头规制与个人赋权制衡

《一般数据保护条例》第 5 条规定了个人数据处理的合法性、合理性、透明性原则，目的限制原则，数据最小化原则，准确性原则，限期保存原则，完整性与保密性原则等六

⑨ *See* Rachel Metz, *Microsoft's Neo-Nazi Sexbot Was a Great Lesson for Makers of AI Assistants*, at https://www.technologyreview.com/2018/03/27/144290/microsofts-neo-nazi-sexbot-was-a-great-lesson-for-makers-of-ai-assistants/ (Last visited on January 15, 2022).

⑩ Lessig 提及的法律规制包括数据赋权（对数据主体）与数据削权（对数据处理者）并对违反者施加责任，即事前与事后规制思路。*See* Laurence Lessig: *Code Version 2.0*, Basic Books Press, 2006, p.223.

项原则，界定了数据控制者对个人数据的存储与处理的权限。如果算法自动决策收集、训练、处理的数据涉及个人数据，则上述原则也适用于算法自动决策中的数据处理，在一定程度上为算法个人数据处理提供了规制基础，避免不当收集数据、使用错误数据或超期限使用数据进行算法决策对个人权利造成侵害。此外，《一般数据保护条例》采取数据主体权利的进路，在序言第 71 条及正文第 12—22 条赋予数据主体获取数据处理相关信息权利、对个人数据的访问权、更正权、删除权（被遗忘权）、限制处理权、数据携带权、一般反对权和反对自动化处理的权利。《一般数据保护条例》通过创制数据权利回应完全自动化处理对个人权利与自由的潜在风险，赋予其面对算法标签、算法操纵等隐形侵害的防御性盾牌，改善其在算法社会中的弱势地位；访问权、解释权规定则试图赋予个人面对算法黑箱及科技企业的信息技术优势以信息平等装备，对抗算法权力的无限扩张。

《个人信息保护法》第 5—8 条规定了合法、正当、必要、诚信、明确、合理、最小化、公开透明等原则，第四章规定个人对关于其个人信息的处理享有知情权与决定权、查阅复制权、可携带权、更正补充权、删除权、解释说明权等（第 44—48 条），在算法规制方面也采用了数据源头规制与数据赋权思路。

2. 个人数据源头规制与赋权方案的缺陷

受制于个人自由主义的价值基底与数据主导的思维路径，当前的立法在算法规制方面存在先天缺陷，在算法侵害防治方面显得有些力不从心。

第一，《一般数据保护条例》和《个人信息保护法》仅适用于利用个人数据的算法自动决策，包括个人直接提供的数据（如通过问卷调查等）、通过观察个人形成的数据（如通过申请收集的住址数据）、通过已建立的个人画像推断的数据。⑪ 如果算法系统处理的数据是无法直接或间接识别到个人的非个人数据 ⑫，如用户使用相关程序或信息服务后产生的无法识别特定个人的数据，或者数据处理者通过数据挖掘技术对原始数据进行匿名化处理及数据清洗、分析、建模后产生的数据，包括搜索引擎记录、电子商务记录、用户使用习惯、潜在用户群等，就无法为上述法律所覆盖，存在法律漏洞。人工智能算法系统往往故意绕开个人数据门槛，精心利用非个人数据对个人或其所属群体进行敏感的推断与预测。鉴于人工智能系统善于根据类别、群体与模式等进行算法画像，欲将算法画像所带来的损害降至最低，设置个人数据门槛可能是不恰当的。⑬ 此外，算法能够分析本身不包含个人数据的数据组，建立数据间的联系，对非个人数据去匿名化，重新追溯至个人，构成对个人数据保护的新的风险。⑭ 通过个人赋权的个人主义路径过分聚焦于具

⑪ *See Article 29 Working Party's Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation*, *supra* note 6, 8.

⑫ 参见程啸：《论大数据时代的个人数据权利》，载《中国社会科学》2018 年第 3 期，第 109 页。

⑬ *See* AI Now Institute, *New York University Submission to the European Commission on "White Paper on AI - A European Approach"*, at https://ainowinstitute.org/ai-now-comments-to-eu-whitepaper-on-ai.pdf, p.4 (Last visited on January 15, 2022).

⑭ *See White Paper On Artificial Intelligence - A European Approach to Excellence and Trust*, at https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf, p.11 (Last visited on January 15, 2022).

体的针对个人的算法侵害，如算法歧视、算法错误对个人平等权、财产权与人身自由权利的侵犯，而对于抽象的、集体性的算法侵害未予关注。⑮

第二，个人赋权模式作为基于个人信息控制权理论构建的应对方案，⑯在算法应用场景存在逻辑悖论，也面临权利行使困境。知情权、访问权、质疑权、纠正权等数字权利行使本就存在难以逾越的信息与技术障碍。⑰这一问题在算法自动化决策中更为凸显：鉴于机器学习算法的动态性、自主性、迭代性，连算法设计者、应用者都难以知晓和控制自动化决策过程，遑论缺乏专业知识的个人，很难期待其主张上述数字权利抵御算法侵害。

从法律社会学角度视之，人工智能引起的经济与社会技术转型带来法律制度变迁，法律制度也积极地作用于这一转型进程，二者交互影响，也改变着法律基本保障措施的内涵与解释，并形塑法律权利与义务的定义及其实现方式。⑱面对人工智能科技带来的人的异化、弱化与边缘化，数字赋权（digital empowerment）模式创设新的权利类型，如数据权、算法权（程序性权利、动态的参与权利），是一种合理的应激性反应。这些针对数字生产活动而构建的新型权利并不与真实世界的现实权利直接对应或关联，对这些数字权利的侵犯无法根据传统责任理论归责与救济。在我国司法实践中，只有当个人信息关联到隐私权、姓名权等民事权利时，个人信息主体才能通过侵权之诉获得权利救济，而在涉及用户画像或其他自动化决策所带来的不利影响的案例中，法院并未支持当事人请求。⑲《一般数据保护条例》和《个人信息保护法》虽然配置了违反上述数字权利规定的责任条款，但仅依靠这种数据合规的行政责任尚不足以应对算法侵害对现实世界的权益妨害，尚须开发侵权法、刑法等传统教义学责任理论，拓展算法责任框架。⑳

（二）算法程序性规制方案及其缺陷

算法决策的不透明性和缺乏监督给个人正当程序权利带来冲击，利用代码进行决策完全不同于传统法律决策时正当程序原则所要求的“人与人之间叙述性的、可理解的交流”，“这种交流不应简化为软件程序与人之间的交流”。㉑有鉴于此，以技术正当程序概念㉒为中心对算法本身的程序性规制与监管方案日益获得关注，理论与实践开始聚焦于技术上与法律上对算法数据训练、设计、应用全流程的动态性管控的可能性。

---

⑮ *See* AI Now Institute , *supra* note 13, 4.

⑯ 参见丁晓东：《基于信任的自动化决策：算法解释权的原理反思与制度重构》，载《中国法学》2022 年第 1 期，第 109 页。

⑰ 参见前注④，丁晓东文，第 146–147 页；张欣：《从算法危机到算法信任：算法治理的多元方案和本土化路径》，载《华东政法大学学报》2019 年第 6 期，第 22 页。

⑱ *See* Julie E. Cohen: *Between Truth and Power - The Legal Constructions of Informational Capitalism*, Oxford University Press, 2019, p.2.

⑲ 参见姚佳：《论个人信息处理者的民事责任》，载《清华法学》2021 年第 3 期，第 47–48 页。

⑳ 也有学者指出，算法滥用等情形目前主要通过监管或数据治理等方式予以规制，民事责任方面尚未见相关适用和突破。参见前注⑲，姚佳文，第 48 页。

㉑ 参见［美］陆凯：《美国算法治理政策与实施进路》，苏苗罕、丁闻语译，载《环球法律评论》2020 年第 3 期，第 7 页。

㉒ 即自动化决策过程也必须满足被决策对象正当程序权利。*See* Danielle Keat Citron, *Technological Due Process*, Washington University Law Review, Vol.85:1249, p.1249–1314 (2008).

1. 算法公开与算法可解释性

为对抗算法黑箱与算法技术权力，算法公开与算法可解释性制度成为增强算法决策透明性、揭开算法面纱的对策性方案。所谓算法公开，是指公开算法的源代码。2016 年，纽约市尝试立法要求政府机构运用算法决策分配福利、维护治安和行政处罚时，公开算法源代码。[23] 算法可解释性要求算法运作逻辑及输入数据与输出决策结论之间的联系能够为人类所理解，包括以结果为基础的解释（outcome based explanation）与以逻辑为基础的解释（logic based explanation）。[24] 算法可解释性被视为保障个人自治性与技术性正当程序权利、避免算法成本外部化的重要制度性工具。[25]《一般数据保护条例》和《个人信息保护法》在个人数据赋权模式下创制的算法解释权也更多是一种程序性权利，作为个体与自动化决策者之间的信任沟通机制，[26] 发挥着技术正当程序的功能。我国《个人信息保护法》第 24 条专门针对自动化决策过程的透明性、可解释性作出规定。《算法规定》超越了既往个人数据保护思维路径，从算法规制维度引入算法可解释性规定，要求算法推荐服务提供者以显著方式告知用户其提供算法推荐服务的情况，并以适当方式公示算法推荐服务的基本原理、目的意图和主要运行机制等（第 16 条）。此外，该规定还配置了退出个性化推荐的权利等算法选择权、用户标签管理权（第 17 条）等程序性权利，在技术正当程序的意义上保障用户的选择权与参与权。

2. 算法审计与算法影响评估制度

算法程序性控制的事前规制还包括算法审计与算法影响评估等制度。算法审计要求算法开发时必须考虑到能够允许第三人监督、批判及审查算法行为，从而有助于提高算法设计质量及在发生错误时及时调整。[27] 算法影响评估由知名人工智能研究机构 AI NOW 提出，是指对算法可能产生的负面影响、受决策影响对象的访问权设置及实现路径、侵害发生后的审查或纠正机制等进行评估，以构建一个联合内部与外部、专业机构与公众、具有实践性的问责框架，促进专业人士与公众对算法的民主参与。[28] 美国国会 2019 年 4 月提出《算法问责法》，要求公司对自动决策系统进行准确性、公正、偏见、歧视、隐私和安全方面的影响评估。[29]我国学者将其视为构建算法信任、促进算法问责的一种算法治理措施，具有事先预期性而非回溯性特质。[30] 欧洲议会研究服务中心 2019 年 4 月发布

㉓ 参见汪庆华：《算法透明的多重维度和算法问责》，载《比较法研究》2020 年第 6 期，第 170 页。

㉔ *See* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, Fordham Law Review, Vol.87:1085, p.1099 (2018).

㉕ 参见前注 ②，苏宇文，第 173 页。

㉖ 参见前注 ⑯，丁晓东文，第 101 页。

㉗ *See* Nicholas Diakopoulos & Sorelle Friedler, *How Hold Algorithm Accountable*, at https://www.technologyreview.com/2016/11/17/155957/how-to-hold-algorithms-accountable/ (Last visited on January 15, 2022).

㉘ *See* Dillon Reisman et al., *Algorithmic Impact Assessment: A Practical Framework for Public Agency Accountability*, at https://ainowinstitute.org/aiareport2018.pdf (Last visited on January 15, 2022).

㉙ *See* Algorithmic-Accountability-Act-2019, at https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf, p.53 (Last visited on January 15, 2022).

㉚ 参见张欣：《算法影响评估制度的构建机理与中国方案》，载《法商研究》2021 年第 2 期，第 104-105 页。

的《算法责任与透明治理框架》采纳了算法影响评估制度。[31]《算法规定》第 8 条也规定，算法推荐服务提供者应当定期审核、评估、验证算法机制机理、模型、数据和应用结果等。

上述制度皆为程序性制度工具，体现了算法风险预防型流程监管的加强趋势。[32] 它们试图建构一种程序性保障措施，以便预防性地将算法决策纳入一个问责性框架。[33] 不可否认，这对规制算法及抵御算法侵害具有重要意义，但也存在以下局限性。

其一，算法公开、算法可解释性等事前规制本身不仅存在侵犯知识产权、引发算计（gaming）、成本大于收益等缺陷，[34] 而且因算法黑箱难题受到技术可行性的制约。某些算法程序性权利的实现也缺乏技术支持。例如，《算法规定》赋予用户标签管理权，要求算法推荐服务提供者向用户提供选择或者删除用于算法推荐服务的针对其个人特征的用户标签的功能，但算法标签不同于个人信息，如何在技术上比照《个人信息保护法》赋予的个人信息删除权支持用户标签的删除、管理权，尚待实践进一步检验。

其二，算法程序性控制方案旨在提供过程性的降低算法损害发生概率、侵害广度与深度的风险预防性措施，既不能确定地有效避免某个具体算法损害的发生，也无法对已然发生的算法侵害提供符合人类理性与正义要求的规范上可感的修复。换言之，算法程序性控制方案仅能在宏观抽象层面降低整体算法侵害的频度、深度与广度，无法在具象微观层面针对具体算法侵害精准施策，提供个案正义意义上的预防与救济。例如，2020 年，美团配送算法对骑手的操纵剥削引发舆论风暴。面对舆论压力，美团公布其配送算法模型并整合多种估算时间模型调整算法，将原先的订单预计送达“时间点”变更为弹性的“时间段”。[35] 但这一改变无济于算法设计不合理导致骑手交通事故伤害甚至死亡事件的权利救济与追责。

三、算法侵害归责难题及解析

人工智能作为一种颠覆性科技，具有不同于传统科技的特性，带来如下归责难题。

（一）复杂 AI 生态与多元参与主体

人工智能涉及复杂的技术生态。算法的开发、应用涉及算法模型开发与训练者，数据收集、控制与处理者，产品制造者，软件所有人等众多参与主体；复杂算法流程往往由众多编程师集体编写，每个编程师可能仅负责其中一部分代码；诸如自动驾驶、护理机器

---

[31] *See A Governance Framework for Algorithmic Accountability and Transparency, European Parliamentary Research Service Scientific Foresight Unit* (STOA), PE 624.262 – April 2019, at https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf, p.53 (Last visited on January 15, 2022).

[32] 参见张凌寒：《〈个人信息保护法（草案）〉中的平台算法问责制及其完善》，载《经贸法律评论》2021 年第 1 期，第 41 页。

[33] *See* Ari Ezra Waldman, *Power, Process, and Automated Decision-Making,* Fordham Law Review, Vol.88:613, p.624 (2019).

[34] 参见沈伟伟：《算法透明原则的迷思——算法规制理论的批判》，载《环球法律评论》2019 年第 6 期，第 24–36 页。

[35] 参见《美团公开配送算法，送达“时间点”渐变“时间段”》，载腾讯网，https://xw.qq.com/amphtml/20210910A09MZL00，2022 年 1 月 15 日访问。

人等人工智能技术具有融合物联网、感知技术、机械制动等多种技术的特征，需要不同领域主体参与。随着数据在不同技术主体间流动，责任也变得具有流动性、模糊性，产生具体错误结果时往往很难查明和界定是哪个主体有过错。[36]

实际上，前人工智能时代也存在多个主体参与的行为致损时无法确定哪个主体的行为是损害发生原因的情形，对此侵权法与刑法有不同的解决方案。侵权法针对数人基于意思联络不法实施共同加害行为或数人实施危险性行为但无法证明谁是加害人的情形，规定了共同侵权责任。《民法典》第1168条规定："二人以上共同实施侵权行为，造成他人损害的，应当承担连带责任。"若存在组织一体性和组织过错时，也可以适用法人侵权责任。针对多人共同引起犯罪结果的归责，刑法上的解决方案包括单位犯罪责任与共同犯罪责任。在多个犯罪行为人具有组织上的一体性场合，刑法通过拟制人格这种法律构造，将损害结果归属于该组织的集体意志与集体行动，追究法人或单位犯罪刑事责任。更有观点主张赋予智能机器人以法律主体资格（电子人格）解决民事责任问题，[37]或类比法人责任追究智能机器人独立刑事责任，以避免智能机器人致损的可罚性漏洞及降低其刑事责任证明的复杂性。[38]如果多个犯罪行为人之间不存在单位组织上的一体性，可以适用共同犯罪责任。若多个参与人之间具有意思上的共同性（即存在犯意沟通），按照"部分行为、整体归责"的原则将结果共同归属于各参与人，即共同犯罪刑事责任。

但上述传统责任方案在算法决策领域的适用具有局限性。其一，就共同侵权责任与共同犯罪刑事责任而言，算法决策各行为人之间往往不存在明确的意思联络。即使承认过失共同侵权或共同过失犯罪，但如上文所述，算法决策作为各技术主体参与的一系列大数据收集、分析与归类的行为，主观方面认定存在较大难度。其二，就人工智能独立责任与法人责任而言，现阶段人工智能独立责任观点面临意志自由与责任资格证立、责任承担等理论及实践难题，即使对责任概念采取规范的、功能性的理解而将人工智能体视为责任主体，也难以发挥责任的预防功能。[39]欧洲议会赋予人工智能电子人格法律地位并使其独立承担损失赔偿责任的方案受到广泛质疑，其《AI民事责任决议》也否定了AI独立责任的模式。通说认为，在可见的未来，算法自动决策引起的损害归责问题仍需要在自然人或法人的责任框架下予以解决。[40]但是，追究法人责任的问题在于，算法模型

---

[36] *See* Iria Giuffrida, *Liability for AI Decision-Making: Some Legal and Ethical Considerations*, Fordham Law Review, Vol.88: 439, p.443 (2019).

[37] *See European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics* (2015/2103 (INL)), at https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html (Last visited on January 15, 2022).

[38] *See* Gerhard Wagner, *Robot, Inc: Personhood for Autonomous Systems*, Fordham Law Review, Vol.88:591, p.600 (2019). 国内赞成独立主体刑事责任的观点，参见江溯：《人工智能作为刑事责任主体：基于刑法哲学的证立》，载《法制与社会发展》2021年第3期，第111-127页。

[39] 参见时方：《人工智能刑事主体地位之否定》，载《法律科学》2018年第6期，第67-75页；李爱君：《人工智能法律行为论》，载《政法论坛》2019年第3期，第176-183页。

[40] *See* Karni Chagal-Feferkorn, *The Reasonable Algorithm, University of Illinois Journal of Law*, Technology & Policy, Vol.1:111, p.115 (2018).

的设计和训练者，数据收集、控制与处理者，产品制造者，软件所有人等技术主体可能分属于不同的法人或单位，并不具有法人责任所要求的组织上的一体性，也往往不会或无法形成针对算法侵害行为的集体意志与行动。

（二）算法黑箱与因果关系难题

与传统的侵害行为相比，算法侵害的因果进程更加冗长、复杂、不透明，导致不仅难以对算法的决策机制作出人类可理解的解释，甚至将侵害结果追溯到具体的行为主体（通常要求在某个输入行为或数据与输出结果之间建立关联）都变得更加困难。

第一，相关性关系取代因果关系。算法挖掘事物间的相关性联系而非因果性联系。20 世纪 90 年代，科学家运用机器学习系统分析大量肺炎患者的数据得出一个令人惊讶的结论：与正常人群相比，有哮喘病史的肺炎患者死于肺炎的风险更低。原因是，有哮喘病史的肺炎患者往往直接入住 ICU，可以得到更集中的治疗与护理。虽然实际上有哮喘病史患者的风险更高，但机器学习将相关性与因果性相混淆，从而得出上述令人费解的结论。[41] 算法决策的技术本质是利用大数据发现统计概率上的相关性关系，而非作为人类理解世界之逻辑基础的因果性关系，[42] 因此，表面上看来，利用传统的人类认知论中的因果关系概念对依赖大数据相关性运作的算法决策及其侵害进行归责仿佛缘木求鱼。

第二，算法黑箱与不透明性。算法黑箱或技术不透明带来因果关系认定与解释难题。关于算法可解释性的通行观点认为，一个充分的解释至少能够说明输入变量的特征如何与预测结果相关，以便人们能够理解如下问题：模型对发放借贷的决策是否会受到申请者是少数族裔的影响？哪些特征在预测中起到最大作用？[43] 算法模型复杂程度与其可解释性之间呈反向相关关系，除代表简单关系的线性算法模型易于解释外，非参数方法如支持向量机与高斯程序则代表更复杂多样的功能因而难以理解。[44] 随机森林等集成学习方法预测结果产生于集合与加权程序，对人类的理解能力提出了挑战。尤其是深度学习神经网络算法，人类几乎没有办法理解其复杂结构中的多层神经网络的权值。[45] 此外，即使要求算法设计者编写具有可解释性的、透明的算法，往往也只能帮助我们理解它做了什么，却无法理解它为什么要这么做。透明性要求或许有助于找出 AI 的不合法行为、瑕疵设计或者不道德的算法，但可能这些并不足以向普通用户解释具体的决策是如何作出的，例如算法为何拒绝借贷申请、自动驾驶系统为何转弯。[46] 近年来，虽然学界与产业界在可解释人工智能（XAI）研究方面取得一定探索性进展（如谷歌的模型卡片机制、IBM 的 AI 实施清单机制等解释方法），但对复杂机器学习算法进行全面、不区分场

---

[41] *See* Mark A. Lemley & Bryan Casey, *supra* note 8, 1336.

[42] 参见维克托·迈尔—舍恩伯格、肯尼思·库克耶：《大数据时代：生活、工作与思维的大变革》，盛杨燕、周涛译，浙江人民出版社 2013 年版，第 67–94 页。

[43] *See* Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"*, at https://arxiv.org/pdf/1606.08813v2.pdf, p.4 (Last visited on January 15, 2022).

[44] *Id*., 4.

[45] *Id*., 4.

[46] *See* Mark A. Lemley & Bryan Casey, *supra* note 8,1365–1366.

景的解释仍是不现实的。[47]

算法黑箱不仅导致可解释性难题，也导致可追溯性难题。在复杂机器学习算法中，将错误的或有偏见的决策结论追溯到某个具体的数据输入或参数存在困难，往往难以证明算法不当考虑了某种信息或对该信息赋予错误的权重而应对该损害负责。《AI 民事责任决议》的决议说明第 H 条指出，算法不透明导致极难判断谁控制 AI 系统的风险以及识别哪个代码或行为、数据输入导致了损害，从而加大了建立行为与损害之间因果联系的难度，也阻碍受害人获得足够的赔偿。

基于法教义学的因果关系原理，可对上述难题进行如下思考。首先，需明确的是，现代归责理论的因果关系概念早已超越了纯粹自然法则意义上的“规律性地引起与被引起的关系”的含义。以刑法上的因果关系为例，假设没有某行为，结果就不会发生，则该行为就是结果发生的原因（所谓必要条件公式）。[48] 虽然大多数情况下因果关系认定都以科学上能够证明的因果律为基础，但在新型损害结果原因无法被科学验证或解释时，只要大致能排除其他因素导致损害，也可以认定存在概括性的因果关系。[49] 在新科技或产品导致损害结果领域，德国学界也有观点倡导概率性、统计学的刑法因果关系概念。[50] 因此，晚近超越存在论层面的、规范的因果关系概念与算法应用中大数据概率性、相关性的联结并不矛盾，认为相关性关系取代因果关系导致算法归责难题的观点仍停留于自然法则的因果论阶段，并不可取。对于具体的算法侵害而言，只要可将错误（或歧视性）决策结论追溯至某个瑕疵行为或数据输入，即可满足必要条件公式意义上的因果关系要求，也就是说若无瑕疵行为或数据输入，就无错误决策结论，则该瑕疵行为或数据输入是错误决策结论的原因。因而，对于算法侵害因果关系认定而言，在 AI 技术细节意义上理解具体的算法决策如何推导，如算法为何拒绝某个借贷申请、自动驾驶系统为何在特定时点转弯等，并非必要。可见，算法可解释性难题不会导致规范意义上的因果关系理论失效，而算法可追溯性难题虽然给因果关系认定带来困扰，但更多地涉及程序法上算法侵害因果关系查明的问题，并未冲击实体法上的因果关系认定及归责。其次，肯定传统责任理论的因果关系概念对于算法侵害归责的适用性，也需要正视算法可追溯性难题带来的因果关系认定障碍。这些障碍对集成学习方法、深度学习神经网络等复杂算法而言往往是无法逾越的。解决之道有二：一是可适用过错责任推定，缓解技术性归责障碍带来的责任真空与受害人救济问题；二是舍弃因果关系要件设定针对违法行为本身的责

---

[47] *See* Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, Nature Machine Intelligence, Vol.1:206, p.206 (2018); 腾讯研究院：《可解释 AI 发展报告 2022：打开算法黑箱的理念与实践》，载微信公众号“腾讯研究院”，2022 年 1 月 18 日上传。

[48] Vgl. Joerg Eisele, in: Schönke/Schröder, Strafgesetzbuch, 30.Aufl. 2019, Vor § § 13 ff, Rn.73.

[49] Vgl. Eric Hilgendorf, Fragen der Kausalität bei Gremienentscheidungen am Beispiel des Lederspray-Urteils, NStZ 12(1994), S.561 ff.

[50] Vgl.Eric Hilgendorf, Der gesetzmäßige Zusammenhang i. S. der modernen Kausallehre, Jura 10(1995), S.514,519 ff.; Andreas Hoyer, Die traditionelle Strafrechtsdogmatik vor neuen Herausforderungen: Probleme der strafrechtlichen Produkthaftung, GA 143(1996), S.160, 168 ff.

任，以此跳过因果联结及可追溯性难题。

（三）故意或过失难题

算法决策是由多元技术主体参与的一系列大数据收集、分析与归类的行为，各参与主体在某个行为阶段具有何种主观方面往往难以认定。某种算法设计、应用的总体目标可能较易于得知。如，证券交易算法程序的总体目的是交易利益最大化，但一旦运用了不透明的黑箱算法，人们就难以得知它如何作出具体交易决策或是否通过市场操纵获取利润最大化，尤其在神经网络算法与支持向量机等复杂算法中，查明主观意图与因果关系更为困难。[51] 美国学者赛伯斯特（Selbst）以大数据挖掘技术的预测性侦察算法为例，阐述了查明算法歧视侵害主观意图的难度。与人所实施的歧视性对待行为（disparate treatment）不同，大数据挖掘与机器学习算法的歧视性影响（disparate impact）往往不存在明显的歧视意图，法律与判例试图将侵害追溯到某个人类行为，即引起歧视的筛选机制的选择之上，但大数据挖掘在某种意义上就是利用机器学习技术在不同人群数据中发现模式以进行特征归类与分析，有些数据本身就客观反映了特定人群的区别性特征及社会历史形成的不平等社会关系。[52] 以违反注意义务为中心的过失认定在算法决策中也非易事，因为人工智能新兴科技领域算法应用往往还未形成实定法或司法判例上的注意义务标准。帕斯奎尔（Pasquale）指出，随着算法决策在线上线下应用的普及，以故意或过失为基础的行为与责任框架将逐渐失色，向以结果为基础的行为责任框架过渡，因此关键的是哪些结果严重到有必要进行干预规制。[53] 我国学者也基于算法不透明、缺乏数字正当程序的分析，发现算法决策应用中以加害人过错为前提的普通侵权责任面临适用困境，阻绝了受侵害方的正常私力救济渠道。[54]

鉴于算法决策中主观过错认定的困难，英美侵权法中的严格责任原则由于不要求行为人的主观过错，仅要求证明损害的因果关系，因而在人工智能算法决策责任的探讨中受到特别关注。我国也有学者指出，鉴于平台主观过错的认定面临“技术中立面纱”与“行为人与责任人相分离”的困境，监管部门对平台事后监管或追责只能借助严格责任或者无过错责任。[55] 我国《民法典》侵权责任编第 4 章产品责任、第 6 章环境污染和生态破坏责任、第 8 章高度危险责任、第 9 章饲养动物损害责任、第 10 章建筑物和物件损害责任等皆采用了严格责任原则。严格责任在欧洲大陆法上也被称为危险责任，[56] 欧洲大陆法系的一些严格责任规定甚至仅要求行为风险的实现或损害可追溯到特定风险

---

51 *See* Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, Harvard Journal of Law & Technology, Vol.31:889, p.907, 920 (2018).

52 *See* Andrew D. Selbst, *Disparate Impact in Big Data Policing*, Georgia Law Review, Vol.52:109, p.124,165 (2017).

53 *See* Frank Pasquale, *Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society*, Ohio State Law Journal, Vol.78:1243, p.1249 (2017).

54 参见唐林垚：《人工智能时代的算法规制：责任分层与义务合规》，载《现代法学》2020 年 1 月，第 198 页；张凌寒：《商业自动化决策的算法解释权研究》，载《法律科学》2018 年第 3 期，第 67 页。

55 参见张凌寒：《网络平台监管的算法问责制构建》，载《东方法学》2021 年第 3 期，第 24 页。

56 参见朱岩：《危险责任的一般条款立法模式研究》，载《中国法学》2009 年第 3 期，第 31–32 页。

群体的行为，例如从事危险活动或运营危险设备的人须对危险的现实化承担责任，即使这些行为无法事先预见，对该风险群体具有管控职责的人也应当对损害负责，如饲主对动物或监护人对精神病人、儿童等特定人的责任。[57] 可见，严格责任不仅松动了传统过错责任原则对主观过错的要求，也根据科技产品的准入与风险许可要求对因果关系进行修正。关于严格责任的实定法规定与判例，无疑提供了应对算法风险及解决侵害归责问题的有益思路。

（四）自主性与失控

人工智能具有自主学习、自主决策与执行决策的能力，此即区别于传统技术的自主性特征。自主性增强了人工智能技术应用效能，但也带来不可预测性与人类控制性减弱甚至失控问题。算法往往基于海量的、动态的数据与参数组合进行决策，编程师无法对所有场景下的决策进行预测，除少数情形需要人对决策结果进行审核外，其他情形下算法将会根据人类没有机会考虑的信息自主作出决策。[58] 人类的有限参与极大弱化甚至架空了人类对算法决策过程的控制，但控制与支配是法律责任理论的一个逻辑基点：人类对行为承担责任的根据在于对该行为具有一定程度的控制。[59] 在侵权法上，过错责任原则中行为的控制自不待言，即使危险责任或严格责任也要求行为人对危险源具有控制力。《AI 民事责任决议》也提倡基于以下责任正义观对高风险 AI 系统造成的损害施加严格责任：创设或管理科技风险或从风险中获利者，应事前采取风险最小化措施，事后对风险现实化造成的损害承担责任（正文说明条款第 8 条）。管理风险、采取风险最小化措施显然须以对风险的控制为前提。就刑事责任而言，至少在支配犯领域，对损害因果流程或构成要件行为进程的支配是归责的重要线索，根据因果关系与客观归责理论这一学界通说，只有行为创设了风险并且风险并非异常地实现——即行为人的行为主导了因果流程时，才能对行为进行归责。[60] 在共犯领域，犯罪行为支配也是正犯责任判断的通说。关于不作为犯罪处罚根据的有力学说“结果原因支配说”，主张对造成结果的原因的支配是作为与不作为共同的归责基础——对造成结果原因具有支配的一方，应对损害结果的发生承担作为或不作为的责任。[61]

自主性带来的失控无疑也对基于控制与支配的归责逻辑造成冲击。《AI 民事责任决议》在回应算法技术特征带来的归责难题时，虽然承认机器学习算法的不透明性与自主性带来可追溯性难题，但试图援引传统教义学责任理论中的控制与支配思想，将 AI 全价值链中创设、管理风险的人视为关键责任人（引言第 7 条）。但这一方案似乎存在矛盾：

---

[57] *See* SWD (2018) 137 final, *Commission Staff Working Document Liability for emerging digital technologies*, at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018SC0137, p.8-9 (Last visited on January 15, 2022).

[58] *See* Karni Chagal-Feferkorn, *supra* note 40, 134-135.

[59] *See* John Martin Fischer & Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge University Press, 1998, p.54, 240.

[60] Vgl. Claus Roxin/Luis Greco, Strafrecht Allgemeiner Teil Band I, 5. Aufl. 2020, § 11, Rn. 44 ff.

[61] Vgl.Schuenemann, Grund und Grenzen der unrechten Unterlassungsdelikte, zugleich ein Beitrag zur strafrechtlichen Methodenlehre, 1971, S.235 ff.

一方面承认复杂 AI 系统的失控，另一方面在责任设置时又假设控制存在，将“管理”风险者与创设风险者均作为关键责任人。

从机器学习算法技术原理上来看，算法黑箱也存在程度区别。一些算法模型、参数或输入数据权重的算法难以被人类理解或控制，即所谓“强黑箱算法”；一些算法具有简单的线性算法模型，或虽复杂，但其模型、参数或输入数据权重仍能够为人类所理解，即所谓“弱黑箱算法”。[62] 因此，可采取一种算法技术区别化的责任方案：对行为人能够控制的“弱黑箱算法”，适用过错或过错推定责任，但为应对过高的因果关系与主观过错查明成本，在无法查明行为人过错时可例外适用严格责任；对行为人无法控制的“强黑箱算法”，则应当跳脱因果支配思维，寻找一种恰当的责任模式解决算法侵害责任问题。

四、构建技术与法律融合视角下的责任框架

上文对传统法律责任框架在人工智能算法领域的适用性分析表明，法教义学责任理论对算法决策新技术特征引致的侵害归责与追责难题并非一筹莫展，提供了不同的责任理论选择。即使在无法追溯引发算法侵害的具体技术主体、因果关系或过错的情形下，传统责任理论在 AI 时代也并非已经死亡，集合式的责任形式（共同责任或法人责任）或严格责任、抽象危险犯责任均可被激活适用并在一定程度上满足算法侵害追责需求。但传统责任理论并非没有局限性，复杂的算法技术生态、算法黑箱、自主性与开放性也为其设定了适用的边界。因此，有必要根据算法技术特征对传统教义学中的责任理论进行调适、拓展，为算法侵害的追责与救济构建一个层级性的责任框架。

（一）传统过错责任原则及过错推定原则

过错责任原则仍然适用于算法侵害。鉴于算法并非客观中立，体现了设计者、应用者的某种价值设定，算法决策机器学习涉及的目标定义、数据准备与训练、特征提取、模型训练与调优等各环节均是计算机编程或数据分析师的行为，[63] 如果能够证明其在上述过程中具有引起算法侵害的故意或过失，传统教义学中的过错责任原则就不应因涉及算法而有所区别。国内学者也指出平台算法的设计部署蕴含着平台的主观意图和价值判断，主张将平台设计部署算法方面的主观过错作为问责依据，并通过算法备案与评估等制度的设置追溯主观过错、固定问责点。[64]《算法规定》第 16 条要求算法推荐服务者公开算法基本原理、目的意图和主要运行机制，也有利于其主观过错的查明。过错责任原

---

[62] *See* Yavar Bathaee, *supra* note 51, 905–906.

[63] *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, University of California, Davis Law Review, Vol.51:653, p.669–698 (2017).

[64] 参见前注 [55]，张凌寒文，第 28 页。但需注意的是，算法设计、部署的主观过错并不完全等同于具体场景下算法决策及其侵害的主观过错，因此对于具体算法侵害仍须结合因果关系等其他归责条件进行具体的归责判断。上述观点似有从完全的结果归责转向完全的主观归责之嫌，仅根据平台设计部署算法的过错认定其对具体结果归责。如，即使滴滴出行将网约顺风车定义为一个“性感的社交场景”，显然也不能将车内发生的强奸罪归责给滴滴平台。

则可以广泛适用于算法歧视、算法错误所造成的算法侵害。

此外，鉴于因果关系证明的困难及算法设计、应用者的信息与技术优势地位，可以适用过错推定原则，要求具有信息或技术优势地位的算法设计者、应用者证明其输入行为或数据对损害不具有因果关系，否则推定这种因果关系存在。例如，《AI 民事责任决议》建议对高风险 AI 之外的损害沿用过错责任制度，并允许运营者通过证明自己遵守注意义务而免责。我国《个人信息保护法》第 69 条第 1 款也规定了过错推定原则："处理个人信息侵害个人信息权益造成损害，个人信息处理者不能证明自己没有过错的，应当承担损害赔偿等侵权责任。"

（二）责任框架的扩展与调适

针对算法的技术特殊性所带来的归责难题，笔者尝试结合传统法教义学资源作如下责任框架的调适与扩展。

1. 非个人化责任：集合式责任

《AI 民事责任决议》提倡，鉴于 AI 系统的不透明性、网络开放性与连接性、自主性，实践中往往很难将 AI 的损害行为追溯至某个特定行为人的行为或设计决策，可借鉴已广泛接受的责任理念逾越这些归责障碍，将 AI 全价值链中创设、维护及控制风险的人即 AI 运营人（Operator）确定为责任主体。所谓运营人，包括：前端运营人（fronted operator），即对 AI 系统的操作运营与功能相关风险具有一定程度的控制，并从运营中获利的自然人或法人；后端运营人（backend operator），即因持续定义技术特征、提供数据或基础支持服务而对 AI 系统的操作运营与功能相关风险具有一定程度控制的自然人或法人（第 3 条）。但问题是，机器学习算法的自主性特征令人类对算法决策过程的控制减弱，找到风险控制方本身就面临技术可行性难题。这启发我们改变将损害追溯至某个特定行为人的思路。传统侵权法上的共同侵权责任以及刑法上的法人犯罪、共同犯罪责任虽然扩大了归责对象的范围、对相关行为人进行整体归责，但也无法完全胜任算法侵害归责问题，存在适用上的局限性。因此，在各行为人之间不存在或无法证明侵害意思沟通而无法适用共同侵权或共同犯罪责任时，以及各技术参与主体之间不具有组织一体性而无法追求法人责任时，有必要创立新的集合责任形式，将数据生产链上的所有行为主体视为具有技术上的一体性，或者将提供某种算法应用而获取商业利益的人视为具有商业上的一体性，[65] 从而要求他们承担集合式的责任。这种集合式的责任可以是故意或过失形式，一般而言应以侵权责任为主，可要求各技术参与主体承担连带赔偿责任。基于刑法的最后手段性与法益保护辅助性原则，[66] 仅在可能造成严重的算法损害时（例如严

---

[65] 欧盟新兴科技与责任专家组提议，认定不同技术参与主体是否构成一个商业与技术单位时可以考虑如下因素：是否对不同技术元素有共同的或相互配合的市场、技术互相依赖与协作的程度以及这些主体之间组合的特殊性与排他性。*See Expert Group on Liability and New Technologies - New Technologies Formation (EG-NTF), Report on Liability for Artificial Intelligence and Other Emerging Digital Technologies (2019)*, at https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/JURI/DV/2020/01-09/AI-report_EN.pdf, p.56. (Last visited on January 15, 2022).

[66] Vgl. Claus Roxin/Luis Greco (Fn. 60), § 2, S. 86 ff.

重的生命健康权、财产权侵害），才比照法人责任设立相应的技术或商业单位刑事责任。

2. 无过错责任：严格责任或危险责任

鉴于算法决策中主观过错认定与结果归责难题，可将产品责任、环境污染责任、高度危险责任以及饲养动物损害责任领域的严格责任或危险责任扩展至人工智能或算法侵害领域。通过严格责任破解算法黑箱带来的主观过错和因果关系难题，似乎是颇有意义的选项。《AI 民事责任决议》主张 AI 风险控制者——AI 运营者（包括前端运营者与后端运营者）对高风险的 AI 应用承担严格责任。该决议提倡针对高风险的自动 AI 系统建立一个普遍的严格责任制度。其第 15 条规定："如果一个 AI 系统的自动化工作蕴含侵害他人的严重风险，以具有随机性、无法被合理预测的方式运作，即可认为具有高风险。判定是否具有高风险应考虑严重风险可能发生的领域及活动的性质、可能造成损害的严重程度、AI 系统应用的方式与产生损害可能性之间的关系等。"2021 年 4 月 21 日，欧盟委员会提交了《关于欧洲议会和理事会制定人工智能协调规则（人工智能法案）及修订联盟部分立法的条例的提案》（以下简称《人工智能法案》），将如下可能对自然人健康、安全造成损害或对基本人权造成负面影响的应用列为高风险 AI 应用：生物识别、关键基础设施运营管理、教育或选举培训、就业及员工管理、获取关键私人服务或公共服务及利益、犯罪及累犯风险分析等执法领域、移民或边境管理、司法和民主进程领域等。[67] 我国可借鉴该方案界定高风险的算法应用场景，对其适用严格责任原则。《算法规定》第 23 条引入算法分级分类安全管理制度，根据算法推荐服务的舆论属性或者社会动员能力、内容类别、用户规模、算法推荐技术处理的数据重要程度、对用户行为的干预程度实施分级分类管理，可依此根据算法对个人或集体权利的侵害类型及程度确定高风险类的算法应用场景，适用严格责任原则。

近来学界对严格责任的态度变得更为审慎：不要求证明行为人对损害具有主观过错即可追究责任，这除了可能遏制人工智能科技创新的积极性，带来寒蝉效应以外，也不利于激励算法设计及应用者遵守合理标准防范损害风险。[68]因此，应限定严格责任的适用。例如，欧盟将严格责任限定于高风险 AI 系统应用场景并根据技术发展及时更新高风险 AI 清单。但这种场景式的限定并未关照算法技术特征，高风险 AI 如果采用欠缺人类控制可能性的复杂机器算法，仅仅因为可能产生损害自然人健康、安全或基本人权的严重后果即对其一体施加严格责任，似有结果责任主义之嫌。因此，笔者提倡采取一种技术限定的进路，即根据算法复杂程度与人类可理解性、可控制性程度，在充分发掘法教义学资源基础上设置不同类型的责任。

3.非因果支配型责任：合规型责任与义务犯

复杂 AI 生态、算法开放性、迭代性、算法黑箱等导致的算法侵害因果关系难题，以及

---

[67] *See* COM (2021) 206 final, 2021/0106(COD), Proposal for a Regulation of the European Parliament and of the Council, Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, Annex III, at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (Last visited on January 15, 2022).

[68] *See* Karni Chagal-Feferkorn, *supra* note 40, 121.

算法自主性导致的失控问题，对依据因果进程归责的支配犯责任造成冲击，即使严格责任也未必能够妥当回应此种冲击与挑战。在英美法上严格责任假定某种控制与可预测性，失去这一假定前提，严格责任适用于复杂机器学习算法造成的侵害就丧失了预防功能与效益基础，因为如果行为人无法预测算法可能得出的结论或造成的后果，便无法据此调整其行为，无法采取预防措施避免损害发生或者精准估算他愿意承受的经济风险。[69] 因此，在复杂的机器学习算法中无差别地适用严格责任并不妥当。而且复杂的不透明的算法更应该被问责，“太复杂而无法解释”不能成为逃避的理由，否则人们就会被鼓励编制复杂算法来逃避责任。[70] 我国也有学者指出，即使无法具体解释算法是如何产生结果的，算法控制者也应对其所使用的算法决策负责。[71]

实际上，在因果支配型归责模式之外，传统教义学的责任模式还存在不要求具体损害结果的责任模式，即仅处罚具有导致损害结果风险的行为（如刑法上的抽象危险犯），或者仅处罚违反法律法规的行为本身（如行政违法行为或不合规行为），以及虽然要求损害结果，但处罚的行为人并非引起该结果或支配该结果的因果进程的人，而是因负有某种义务而对损害承担责任的人。

（1）对于强黑箱算法造成具体法益或权利损害的具体算法损害，借鉴德国刑法教义学中的义务犯理论与私法上的信息信义与公共义务理论，要求设计、应用算法的大型平台或智能产品的生产商、运营商对算法侵害风险承担损害防止义务或对用户承担保护义务，即使不能对算法决策施加完全的支配或控制，但仍应在认识到风险可能现实化时承担损害防止的责任。但鉴于义务性责任缺乏对结果的控制与支配，因而应配置更轻的责任。该方案可在刑法教义学中找到理论资源。刑法归责体系存在支配犯与义务犯的分野。所谓支配犯，是指大多数犯罪如故意杀人罪、盗窃罪等，刑法将这类社会不可接受的行为作为构成要件行为详细描述并禁止，行为人实施该行为、支配因果流程引发损害结果，因而支配成为该类犯罪结果归责的重要线索。所谓义务犯，是因为违反某种社会角色（父母、监护人、官员等）的行动要求（社会角色期待或制度性义务）而造成损害或未阻止损害的犯罪，对于这类犯罪的结果归责，重要的不是对现实损害因果流程的支配，而是与损害结果或被害人具有某种特定的制度性的关系。[72] 社会科技进步形塑新型社会关系，也构造了新的社会角色及义务。算法社会催生了一种信息权力关系，即设计、应用算法的公私部门因巨大的信息优势拥有对被决策对象的权力，因而对其负有保护和损害防止的义务。[73] 具体而言，算法设计、应用者对其客户、终端用户、消费者具有信息信义义

---

[69] *See* Yavar Bathaee, *supra* note 51, 894, 931.

[70] *See* Kirsten Martin, *Ethical Implications and Accountability of Algorithms*, at https://link.springer.com/content/pdf/10.1007/s10551-018-3921-3.pdf (Last visited on January 15, 2022).

[71] 参见解正山：《算法决策规制——以算法“解释权”为中心》，载《现代法学》2020年第1期，第190页。

[72] Vgl. Claus Roxin, Strafrecht Allgemeiner Teil Band II, 1. Aufl.2003, § 25, Rn.267; Guenter Jakobs, Strafrecht Allgemeiner Teil, 2.Aufl.1993, 7.Abschn.70 f, 21. Abschn.16 f, 47 ff.

[73] *See* Jack M.Balkin, *supra* note 5, 1226-1227; Frank Pasquale, *supra* note 53, 1244-1246.

务（Information fiduciary duty），对其所托付的信息数据泄露承担责任，并不得利用信息数据对其造成损害；算法设计、应用者对用户和客户以外的公众具有公共义务（public duty），这种公共义务的核心是避免算法应用的成本或损害外部化，采取措施避免损害或提供救济。这种新型的社会角色关系也拓展了刑法教义学中义务犯归责思想适用的疆域，以解决因人类因果性支配降低甚至消失而带来的归责难题，填补强黑箱算法损害责任真空。欧盟《数字服务法（草案）》《数字交易法（草案）》与我国《个人信息保护法》均积极回应数字经济挑战，引入大型网络平台的“守门人”特别义务。[74] 更广泛的义务也意味着更广泛的责任范围。《算法规定》第 7 条规定：“算法推荐服务提供者应当落实算法安全主体责任，建立健全算法机制机理审核、科技伦理审查、用户注册、信息发布审核、数据安全和个人信息保护、反电信网络诈骗、安全评估监测、安全事件应急处置等管理制度和技术措施，制定并公开算法推荐服务相关规则，配备与算法推荐服务规模相适应的专业人员和技术支撑。”这一立法与监管趋势有助于逐步完善算法应用者的行为规范与注意义务，也为算法的事后规制与归责埋下实定法的伏笔。

可见，即使算法设计者、应用者无法对算法进行传统意义上的因果支配，但根据人类社会的法律文化与前述责任理论，并不意味着其可以完全置身事外。例如，就第三人介入的算法滥用损害而言，设计发布聊天机器人 Tay 的微软公司应当预见到它可能会在与用户的交流中自主学习，表达传播种族主义和色情的信息内容。“鉴于人工智能算法系统伦理、数据伦理的文献已经汗牛充栋、深入人心并且唾手可得，研究者不应以颠覆性科技试验的外衣为其掩盖、推脱责任。”[75]

（2）针对算法标签、算法操纵等抽象的算法妨害，如上所述，算法妨害指企业或政府大规模应用算法给个人或集体带来就业、信用、贷款及社会地位方面的不利影响，增加了其被监控、操纵、歧视与社会排挤的可能性。算法的应用虽然提高了效率，但也带来了成本外部化，类似于公共妨害中的环境污染行为。环境损害中因果关系认定难题，促使环境污染犯罪构成要件逐步放松对结果的要求。2011 年《刑法修正案（八）》将刑法第 338 条重大环境污染事故罪修改为污染环境罪，取消重大环境事故这一构成要件结果要求。司法解释也规定存在超过一定标准排放污染物的行为即构成犯罪，本罪由结果犯转变为行为犯或抽象危险犯。[76] 与环境污染类似，算法妨害也具有累积性和群体性效应，即对个人可能造成的侵害是微小甚至不可感的，但是累积起来却对社会和集体造成可观的侵害。未来，当算法妨害严重或聚集达到足以启动国家刑罚权时，即在满足需罚性前提

---

[74] 参见张新宝：《互联网生态“守门人”个人信息保护特别义务设置研究》，载《比较法研究》2021 年第 3 期，第 16–24 页。

[75] Frank Pasquale, *supra* note 53, 1248.

[76] 参见 2016 年最高人民法院、最高人民检察院《关于办理环境污染刑事案件适用法律若干问题的解释》（法释〔2016〕29 号）第 1 条。

下，可以通过设立相应的抽象危险犯对上述算法妨害行为加以禁止，[77]施加刑事责任对损害集体性法益的行为予以一般预防与特殊预防。

因而，算法妨害的责任设定不以风险的现实化为前提，避免了算法可解释性与可追溯性问题带来的归责难题，将算法标签、算法归化、算法操纵规定为行政违法行为，或类似于公共妨害的侵权行为，追究算法设计者或应用者行政责任或侵权责任。在大型平台或公共机构大规模实施上述行为可能造成广泛、严重的算法侵害时，甚至可将其规定为抽象危险犯，追究刑事责任。《算法规定》第 10 条禁止将违法和不良信息关键词记入用户兴趣点或者作为用户标签并据以推送信息。第 14 条禁止利用算法虚假注册账号、非法交易账号、操纵用户账号或者虚假点赞、评论、转发以及利用算法屏蔽信息、过度推荐、操纵榜单或者检索结果排序、控制热搜或者精选等干预信息呈现，并配置相应合规行政责任条款。这是以联合发布部门规章的形式对算法标签与算法操纵侵害进行行政监管的尝试。未来我国可结合算法监管实践，在相应部门法或统一的“算法法”中将算法妨害行为规定为行政违法行为、侵权行为甚或刑法构成要件，配置行政、侵权甚至刑事责任，形成具有独创性的中国算法事后规制的责任方案。考虑到民事侵权损害赔偿制度难以获得有效救济及刑罚的“最后手段性”，可设定类似守门人特别义务的义务型责任（合规责任），构建算法设计者、应用者（往往是大型在线平台）行政责任为主、民事与刑事责任为辅（义务犯）的多元协同的法律责任机制。[78]

五、结　语

算法决策嵌入社会，带来系统的、多维的算法侵害，包括抽象的算法妨害与算法歧视、算法错误等具体的算法损害。赋权模式与算法可解释性、算法可问责性、算法影响力评估、算法审计等程序性规制方案，无疑为防治算法侵害提供了重要的制度性工具，但难以对算法侵害进行有针对性的归责、追责及救济，因而有必要穿越技术性规制话语的喧嚣，寻回传统法教义学中的责任语词。本文根据既有教义学理论资源与现阶段人工智能技术发展状况，进行归责与责任的开放性探讨，联结算法事前与事后规制版图，以此开启人工智能法律图景的完整性建构实验。

网络法时代，人们逐渐发现并用法律理论发展证明“网络并非法外之地”。人工智能时代，人们也将证明算法并非责任的真空地带。算法新兴技术特征带来的归责难题并

---

[77] 欧盟委员会《人工智能法案》第 5 条将造成或可能造成身体或心理损害的以个人无法查知的方式严重扭曲人类的行为，或利用特定人群的年龄、身体或精神残疾等任何脆弱性实质性地扭曲与该人群有关的行为导致或可能导致该人或他人的身体或心理伤害；公共机构利用个人或社会行为特征进行与收集数据背景不相关的或与社会行为不成比例的可能导致特定自然人或群体不利或不公平待遇的社会评分；除了非常必要的犯罪侦查（包括寻找丢失儿童）、预防恐怖袭击等以外的在公共场所为执法目的而实时远程生物特征识别等列为不可接受的 AI 风险，禁止相关应用但并未科以相应的责任，如若配置针对性的事后规制措施，则正是此处所论的公共妨害或抽象危险犯的责任适例。

[78] 参见前注 [74]，张新宝文，第 23 页。

未宣告人工智能法律责任的死亡，应检验传统责任理论在人工智能算法场景中的适用性与拓展可能性，根据算法技术特征构建一个层级式的人工智能算法责任框架，积极发挥责任制度损害补偿与预防的功能，形成应对算法侵害事前规制与事后规制的完整闭环：对于产生具体侵害结果的算法损害归责而言，分别设置因果支配型的责任与非因果支配型的义务型责任。前者是指针对具有理解与控制技术可能性的算法适用过错责任与严格责任，后者是指针对不具有理解与控制技术可能性的算法适用义务型责任，为算法设计者与应用者设置避免算法侵害风险现实化的义务或对用户、社会公众法益的保护义务及相应责任；对算法妨害而言，因无具体权利侵害结果而不存在结果归责问题，仅需对该算法妨害行为加以禁止，设置相应行政违法行为，或设置类似侵权法的公共妨害甚或刑法上的抽象危险犯进行追责。

Abstract: The technical characteristics of algorithmic decision-making, such as subject pluralism, openness, opacity, and autonomy, pose challenges to the attribution of algorithmic harm. It is necessary to construct a hierarchical algorithmic liability framework based on the testing and expansion of traditional tort law and criminal law liability framework, combined with the technical characteristics of the algorithm. For the attribution of the algorithmic harm that causes concrete harm outcome, causal dominant liability and non-causal dominant duty liability or duty-violating liability should be set up respectively. The former refers to the application of fault liability and strict liability to the algorithm which is explainable and controllable according to its technological characteristics, and the latter refers to the application of duty liability to the algorithm which is unexplainable and uncontrollable according to its technological characteristics, which sets the duty to avoid the realization of algorithm harm risks or the duty to protect the legal interests of users and the public and corresponding liabilities for algorithm designers and deployers. For the algorithmic nuisance, because there is no specific result of right infringement, there is no problem of result attribution, and it is only necessary to prohibit the algorithm nuisance, set up corresponding administrative delicts, or set up the public nuisance similar to tort law or even abstract dangerous offense in criminal law to be held accountable.

（责任编辑：强梅梅）