·人工智能与未来社会(二十六)·

立法还是释法：AI 数据投毒的治理纠偏

周瑞珏[1]　赵精武[2]

【内容摘要】 训练数据作为人工智能技术创新的关键要素，正面临着以降低训练数据质量为目标的数据投毒等新型安全威胁。数据投毒通过各种方式使得训练数据掺杂受污染数据，诱导人工智能输出错误有害信息。《生成式人工智能服务管理暂行办法》第 7 条虽规定服务提供者应当采取有效措施提高训练数据质量，但这种笼统性的规定显然无法适配复杂多变的数据投毒治理实践。现有研究大多主张通过单独立法或者增设专门条款方式细化数据质量保障义务内容，而忽视了数据质量的评价指标不仅包括技术标准，还包括训练数据是否符合个体的业务需求。因此，数据投毒治理框架的建构应当以“释法”为主，基于“数据质量标准 + 数据质量技术工具 + 数据质量管理流程”的数据质量保障范式，明确个体层面数据质量保障义务的履行标准，优化数据质量保障规则与其他人工智能技术治理规则的内容衔接方式。

【关 键 词】 数据投毒　数据质量　数据质量保障义务　数据质量管理流程

【作　　者】 1 周瑞珏，对外经济贸易大学数字经济与法律创新研究中心研究员。（北京 100029）
2 赵精武，北京航空航天大学网络空间国际治理研究基地副主任、副教授。（北京 100191）

【基金项目】 国家社科基金后期资助项目“人工智能法的制度议题与多元理论”(24FFXB014)

为了更好地促进人工智能技术创新，以数据、算力和算法三大核心创新要素为中心的技术治理主张层出不穷。其中，数据要素决定了人工智能如何“认识”和“理解”现实世界，同时也决定了人工智能产品及服务的功能水平上限。为了更好地争夺人工智能市场优势地位和技术创新先机，各大企业将数据资源获取能力的提升作为关键业务环节。然而，在企业追求数据资源又多又好的过程中，数据投毒行为的出现使得“优质数据驱动人工智能创新”的发展目标受到挑战。此前，3·15

晚会相关报道以及新华社都注意到此乱象。“国家安全部”微信公众号指出：“这种通过恶意数据污染 AI 模型的行为，不仅扰乱商业秩序、影响信息传播，更会危害国家安全。”[①]从本质上来看，所谓的数据投毒，是指攻击者通过向人工智能的训练数据恶意植入虚假、恶意或者有偏见的数据样本，从而在模型预训练、微调等环节“污染”大模型的训练过程，导致大模型在输出结果环节呈现不可靠输出的问题。该类行为的危害性不仅仅表现为对人工智能研发进程的破坏、干扰，还会导致违法不良信息的批量化生成，甚至对社会公共利益、国家安全产生严重威胁。在特定场景下，数据投毒行为还可能造成其他权益侵害结果，如在医疗诊断辅助场景下，数据投毒可能对患者的生命健康产生严重威胁。

①《AI“投毒”，危害不容小觑!》，“国家安全部”微信公众号，访问日期：2026 年 4 月 21 日。

尽管《生成式人工智能服务管理暂行办法》在第 7 条明确提及“采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性”，但数据投毒具有技术隐蔽性、损害结果发生延时性等特征，这使得概括性的义务规范难以有效预防和管控数据投毒风险。在一般认知下，鉴于数据投毒作用于训练数据的采集和处理环节，理论上仅需要求义务主体选择来源合法的训练数据、持续强化训练数据清洗加工，似乎就能够解决数据投毒问题。然而，数据投毒作为一种新型人工智能安全威胁，其攻击方式与人工智能技术创新呈现出同步发展状态，“控制好训练数据就能预防数据投毒”的治理逻辑已然无法适应数据投毒多样化的攻击路径。虽然数据清洗、数据验证等技术手段在一定程度上能够保障训练数据的准确性，但难以识别和筛除精心伪装的有毒数据。因此，现阶段数据投毒治理所面临的核心法律问题是，诸如第 7 条等针对训练数据的笼统性义务规范已然无法适配数据投毒行为作用环节多元化、影响结果扩大化的发展趋势。不过，鉴于现行立法已经规定了训练数据相关的法定义务，单独针对数据投毒制定专门立法显然缺乏可操作性，故而该问题的解决仍然需要回归法律解释层面，阐明第 7 条等笼统性义务规范的理论基础和具体内容，进而明确义务主体在预防数据投毒风险方面的义务履行标准。

数据投毒的治理现状及其局限

数据投毒问题早已有之，只不过在传统场景下通过数据清洗等技术措施能够有效解决。但在人工智能领域，传统的数据投毒治理方式难以起效。因此，在以数据投毒作为治理对象时，首先应当明确该类行为的基本特征和表现形态，并结合现有的治理理论和治理实践，厘清当下人工智能数据投毒治理的误区。

（一）数据投毒的基本模态与治理现状

实践中，最常见的数据投毒方式便是在网络上大量发布存在社会偏见、虚假事实等的网络信息，一旦研发者通过网络爬虫等方式抓取网络中的公开数据，这些错误的网络信息很容易对大模型的训练结果产生严重负面影响。由于当下的大模型难以直接处理自然语言，通常会将自然语言转变为向量进行数学计算。这被称为“词嵌入”（word embedding）过程，其使得大模型能够在数理层面理解自然语言的基本内涵。简言之，诸如“合同法”“意思自治”“意思表示”等词语因其语义相近，在数理层面呈现各个词嵌入向量趋同，能够让大模型基于一种“上下文”的逻辑提供较为准确的输出结果。故而数据投毒活动将是依赖词嵌入技术的大模型作为主要攻击对象，[②]通过刻意增加相关词组同时出现的频率，对大模型输出的目标结果进行人为的间接诱导。有学者将数据投毒常见的攻击方式分类为“有针对性的攻击”“无针对性的攻击”“标签投毒 / 后门攻击”“训

② Laura Verde, Fiammetta Marulli,Stefano Marrone, “Exploring the Impact of Data Poisoning Attacks on Machine Learning Model Reliability,” *Procedia Computer Science,* vol.192, 2021.

练数据投毒”“模型逆向攻击”，以及“隐蔽攻击”等几类。[①]但需要注意的是，虽然数据投毒主要针对的是训练数据，但攻击方式并不固定。例如，“标签投毒”需要直接对训练数据进行恶意标注，故而行为方式多是以内部员工恶意标注或者利用安全漏洞进行数据污染为前提；“隐蔽攻击”则是以隐蔽方式对人工智能信息系统进行恶意干预，这类数据投毒的作用方式并不以模型训练环节为限，还包括在用户毫不知情的情况下诱导大模型输出错误结果。

从数据投毒的现状而言，尽管已公开的数据投毒实践案例相对较少，但无法否认该类安全事件的客观存在以及对大模型的安全威胁。例如，儿童手表的人工智能软件因训练数据遭受污染，导致其输出否定中国发明创造等不良网络信息内容，严重误导未成年人；某 AI 软件同样因数据污染而将某地交警业已注销的自媒体账号与一起交通事故强制关联。为此，国家安全机关曾专门发布提示称应警惕人工智能数据投毒，“当训练数据集中仅有 0.01% 的虚假文本时，模型输出有害内容会增加 11.2%”。[②]随着人工智能技术的普及和应用，数据投毒攻击会因相关场景而延伸出不同程度的安全风险：在金融领域，遭受数据投毒的人工智能可能炮制虚假股票信息并影响股价，形成新型市场操纵风险；在国家安全领域，数据投毒则可能诱发社会公众恐慌情绪。

在法律规范层面，我国在建构数据投毒治理体系时并未设置直接且具体的法律条款，而是以网络信息内容治理为内核，强调保障人工智能生成内容的合法性和合理性。纵观现行立法，与数据投毒治理相关的法律规范主要包括三类：第一类是训练数据安全保障条款，如我国《网络数据安全管理条例》第 19 条明确要求，提供生成式人工智能服务的网络数据处理者应加强对训练数据和训练数据处理活动的安全管理，只不过这里的“安全管理”更偏向宽泛意义上的网络数据安全风险。第二类是训练数据质量条款，如《生成式人工智能服务管理暂行办法》第 7 条明确要求义务主体应当增强训练数据的真实性、准确性、客观性、多样性。第三类则是算法安全管理条款，这类条款的核心逻辑是通过规范算法技术应用、训练数据处理等技术环节，预防数据投毒可能导致的各类有害网络信息生成。又如《生成式人工智能服务管理暂行办法》第 4 条明确要求义务主体在训练数据选择、模型生成和优化、提供服务等环节，采取有效措施防止生成式人工智能输出歧视性信息。同时，该办法第 19 条还明确义务主体应当配合有关主管部门就训练数据来源、规模、类型、标注规则、算法机制机理等予以说明。结合《国务院 2026 年度立法工作计划》将立法目标表述调整为“加快推进人工智能健康综合性立法工作”来看，这种“立法空白”实际上是为了预留足够的法律解释和适用空间。因为数据投毒的行为方式在随着人工智能技术的快速发展而同步变化，在现有立法已经足以回应人工智能绝大多数安全风险的情况下，专门针对数据投毒进行单行立法或者增设专门条款都不具有可行性。更何况，2025 年 11 月正式施行的国家标准《网络安全技术　生成式人工智能预训练和优化训练数据安全规范》（GB/T 45652-2025）也提出了在预训练、优化训练、内容生成等环节识别和筛选数据投毒的相关技术标准。此时，继续坚持“通过专门立法推动人工智能技术治理”这一粗放式的治理逻辑，只会导致相关立法无法适应数据投毒技术的快速变化。

（二）数据投毒的治理误区

当下的数据投毒治理实践被部分学者认为存在“规制不足”“针对性不强”“治理方式滞后”等问题，故而相关研究结论呈现精细化、场景化的特点。具体而言，现有研究视角和研究结论大致可分为三类。第一类是风险治理论，即遵循风险分级分类治理的逻辑，结合数据投毒风险的具体特征、损害结果等要素建构具体的数据投毒治理机制。例如，按照数据投毒对人工智能生态构成的影响方式，可划分出其对基础模型层、知识生态层、社会系统层、国家安全层的影响，进而建构全链条技术防

① Frank Hartle III, Steve Mancini, Emily Kerry, “Data Poisoning 2018–2025: A Systematic Review of Risks, Impacts, and Mitigation Challenges,” *Issues in Information Systems*, vol.25, no.4, 2025, pp.433-434.

②《0.01% 虚拟训练文本可致有害内容增加 11.2% 警惕人工智能数据投毒》，央广网，访问日期：2026 年 2 月 18 日。

御机制、全生命周期的数据质量管理制度等治理机制。[①]第二类是刑事犯罪论，即从刑法层面剖析数据投毒的罪名认定。不过，这类研究对于数据投毒也存在显著分歧，围绕数据投毒究竟是否构成非法经营罪、破坏生产经营罪、破坏计算机信息系统罪等有较大争议。[②]也有学者从实际结果考量，认为数据投毒会侵害数据的可用性和完整性，导致计算机系统不能正常运行，故应采用“干扰行为”这一要件的解释路径。[③]第三类则是广义数据治理论，这类研究鉴于数据投毒的作用对象是训练数据，进而在宽泛意义上的数据安全层面讨论数据投毒的治理方式。部分学者将训练数据投毒与训练数据不当收集、代表性不足、泄露等风险一并视为常见的训练数据安全风险，并提出在管制性监管与激励性监管并重的治理模式基础上，建构技术、管理制度和责任三位一体的治理机制。[④]

前述研究在不同层面解释了数据投毒的治理思路，但总体上存在一个突出问题未能解决：在现行立法已经对训练数据安全作出规定的前提下，相关研究仍然在“强化训练数据安全保障”等层面论证数据投毒治理机制，反而使得数据投毒治理这一研究议题缺乏实践价值。总体而言，现有研究之所以未能圆满解释该问题，是因为治理理念存在偏差。数据投毒作为一种新型人工智能安全威胁，其治理方式应当是以技术治理为主，法律、市场等其他治理模式为辅。现有研究未能突破“保障训练数据安全或者质量”等笼统论述的原因是下意识地将法律治理视为唯一或者主要治理工具，进而相关研究结论自然倾向于提出数据安全的一般性治理措施。但是，法律对数据投毒的治理效果有限，且多以事前风险预防为主，主张建构相应的数据投毒法律规范必然无法完成从笼统治理方向到具体治理机制的转换。进一步而言，现有研究所存在的突出问题本质上也是“数据投毒治理是否属于法学研究议题”的另一种表现方式，其混淆了不同治理工具在应对数据投毒风险中的差异化作用。

数据投毒治理是一项综合性技术治理议题，其综合性表现为以技术治理为主，其他治理工具为辅。探讨数据投毒的基本特征、行为方式以及技术原理的目的，是为了更好地解构数据投毒风险的形成过程，进而确认技术、法律、市场等多种治理工具在整个治理框架内的体系定位和治理目标。回顾我国网络安全立法进程，不难发现相关研究同样经历了从“否定网络安全治理属于法学研究议题”到“普遍认可法律应当全面介入网络安全治理实践”的观念变迁。所以，对“数据投毒治理是否属于法学研究议题”这一问题的回答，需要澄清技术治理与法律治理的目标差异。从治理效果来看，通过技术更新、提升安全技术措施等方式能够直接有效地解决数据投毒问题。但是，在现有技术无法根治这类风险的情况下，则需要其他治理工具补足相应的治理空缺。立法论看似能够借助增设专门条款凸显数据投毒治理的针对性，但问题是，面向数据投毒所增补的法律条款既不足以支撑“单独立法”的条款数量，也不足以支撑修法的必要性。在刑法层面，确定数据投毒的罪名适用及其刑事责任能够起到事前威慑、事后惩治的治理效果。然而，数据投毒治理需求主要表现为尽可能在事前阶段将相关风险控制在最低水平，这决定了法律治理目标并不单纯局限于事后的法律责任认定，而是尽可能督促相关义务主体采取一切合理必要措施预防数据投毒风险的发生。至于数据投毒所产生的各类风险层次，本质上不过是输出有害信息内容的不同表现方式，故而在论及数据投毒治理机制建构时，明确相应的治理理论和义务体系建构更显迫切。

以数据质量保障为导向的数据投毒治理理论

数据投毒作为一项新型人工智能安全威胁，无论从风险形态，还是从损害结果来看，均缺乏单独治理的必要性和特殊性。数据投毒治理的讨论应当优先解决其在整个人工智能安全治理框架

① 参见翟岩、李小波：《人工智能安全视域下“数据投毒”的内涵特征、层级风险与治理路径》，《河南社会科学》2025年第11期。

② 参见李怀胜、曹智：《涉人工智能数据投毒行为的刑法应对》，《数字法治》2025年第5期；徐蕴杰：《数据投毒的罪名适用与刑事归责》，《天府新论》2025年第4期。

③ 参见孙本雄、林愉杰：《数据投毒行为的刑法规制困境与纾解思路》，《昆明理工大学学报》（社会科学版）2025年第4期。

④ 参见孙清白：《论人工智能大模型训练数据风险治理的规范构建》，《电子政务》2024年第12期。

中的体系定位，避免以“新瓶装旧酒”的方式重复论证网络安全或数据安全风险的治理方式。

（一）数据投毒治理的起点

现有研究大多遵循“风险特征—风险类型 / 损害结果—治理机制”这一论证逻辑探讨数据投毒的治理方式，但淡化了对数据投毒治理相关理论的探讨。这种论证倾向必然导致数据投毒治理的相关结论以增补法律规范为主，即便设想了诸多较为具体的治理建议，但客观上仍缺乏可操作性，这里的问题是治理逻辑和治理理论的缺失难以有效整合这些治理机制的内在建构逻辑。诚然，在面向数据投毒这类具体的安全风险时，论证相关治理理论有纸上谈兵之嫌，但需要指出的是，治理逻辑和治理理论的阐明直接关系到治理模式和治理机制的选择。事实上，现有研究或多或少已经意识到这一点，在论证过程中试图先行论证数据投毒治理的范畴归类，或是训练数据安全保障，又或是输出内容的风险分层次治理，只是未能建构其治理逻辑、治理理论与治理方式之间的逻辑衔接。换言之，在论证数据投毒治理理论之前，首先需要明确的是数据投毒治理议题应当归并至哪一类治理议题。从作用对象来看，数据投毒是以污染、破坏训练数据完整性、准确性、真实性为目标，相关的治理措施应当围绕训练数据安全保障来建构，故而应当将数据投毒治理议题归并至训练数据安全保障范畴。从作用结果来看，数据投毒导致的损害结果是人工智能技术应用功能减损，输出各种有害信息，相关的治理措施更应当围绕人工智能生成内容管控予以保障。从作用机制来看，数据投毒是以干扰算法优化迭代为重心，输出诸如职业歧视、贬损特殊人群等有害信息，属于算法偏见问题在人工智能领域的延伸，该议题更适合归并至算法安全治理范畴。

数据投毒治理实践涉及算法、训练数据以及大模型生成内容等多个治理对象，若从作用过程来看，数据投毒的侵害客体始终以训练数据为中心，而且“训练数据失真是人工智能幻觉的源头”，[①] 故而更应当将数据投毒治理机制归并至训练数据安全保障范畴。数据投毒的行为特征是通过污染数据等方式干扰大模型算法优化迭代过程，并致使大模型输出有害信息，或者在数据中植入触发器，一旦大模型得到特定输入指令即会触发指定的输出结果。数据投毒所导致的损害结果仅有严重程度之区分，在法律性质层面并无差异，即无法按照预期稳定可靠输出相关结果。因此，如若将之归并至人工智能生成内容管控议题，则难以凸显数据投毒治理所需要回应的训练数据安全保障问题。并且，依照人工智能生成内容管控的治理逻辑，数据投毒的治理模式会被推导至研发者、服务提供者应当采取诸如屏蔽、中断等必要措施预防输出有害内容，这与现行立法中有关网络信息内容治理的相关条款并无实质性差异。例如，《生成式人工智能服务管理暂行办法》第 4 条、第 14 条均提及禁止人工智能生成违法有害信息；《网络信息内容生态治理规定》第 2 章和第 4 章则明确了网络信息内容生产者、网络信息内容服务平台对网络信息内容进行治理的法定义务，形成了从人工智能生成内容到用户使用生成内容、平台管理生成内容多个环节的连贯式治理架构。如此一来，面向数据投毒设置相应治理机制的观点反而缺乏必要性，因为现有法律规范足以应对这类安全威胁。类似地，数据投毒对算法功能的影响也是借由输出内容予以评判，故而不适宜将数据投毒治理简单归并至算法安全治理议题。

在训练数据安全保障体系下，数据投毒的治理逻辑则需要围绕“如何保障训练数据安全可信”这一基础问题展开。一方面，数据投毒对训练数据的影响方式并非传统观念中的数据安全风险，其之所以会产生研究议题范畴定位模糊的问题，是因为数据投毒并没有直接泄露、销毁训练数据，而是依照“训练数据—预训练、优化训练、微调—算法—输出内容”这一系列技术环节产生负面影响。在既有的训练数据安全制度框架中，“安全”更多是指训练数据能够保持私密性、完整性，

① 解志勇、吴梦玉：《人工智能基础模型提供者的系统安全保护义务》，《比较法研究》2026 年第 1 期。

在技术安全措施和内部管理措施的双重保护下有效预防数据泄露、数据毁损等风险。但是，数据投毒对训练数据的损害形式更多是以掺杂劣质数据、恶意数据为主，难以用狭义层面的数据安全予以涵盖。不过，若是从广义的数据安全层面理解，数据投毒所导致的训练数据安全风险更应当被描述为数据质量安全风险，即训练数据的准确性、真实性受到挑战。因此，数据投毒的治理逻辑应当以保障训练数据质量为起点。另一方面，数据投毒具有隐蔽性、短期迭代性、长期影响性等特征，相应的治理逻辑应当以综合性治理为核心。这里的“综合性”并非刻意强调采用多种治理工具，而是需要注意到投毒方式包括直接污染训练数据、网络上批量发布错误网络信息间接污染训练数据等，相应的治理措施应当与人工智能生成内容管控、网络信息内容生态治理等其他治理机制保持衔接，进而达成在多个技术环节综合管控数据投毒风险的治理效果。

（二）数据投毒治理的应然逻辑

数据投毒治理的逻辑起点应当是保障训练质量，而非后续的输出内容管控。在作用逻辑上，“数据投毒可能导致安全驾驶决策失误、医疗诊断结果错误等损害结果”的表述并无问题，但如若表述为“安全驾驶决策失误、医疗诊断结果错误等损害结果是因为数据投毒所导致的”，却存在一定逻辑偏差。因为现有技术水平无法保障人工智能产品和服务能够达到绝对安全可靠的程度，并且，安全驾驶决策失误、医疗诊断结果错误等损害结果可能是多种原因共同导致的。在治理逻辑层面，更应当将治理重心放在训练数据质量保障环节，至于因训练数据不准确等原因所产生的后续损害结果则可以放置于人工智能技术治理框架的其他部分。值得注意的是，强调数据投毒的治理逻辑应当是数据质量保障并非空中楼阁，而在现行立法和产业政策文件中均有体现，只不过尚未形成体系化的数据质量保障治理范式。更确切地说，数据质量保障这一治理目标面向的是整个人工智能训练数据治理体系，内在逻辑上并不表现为为了治理数据投毒而重新创设一个新的治理目标。工信部人工智能标准化技术委员会在 2025 年发布的《工业和信息化领域人工智能安全治理标准体系建设指南（2025 版）》，建构了包含应用安全、算法模型安全、数据安全、网络安全、基础安全等在内的人工智能安全治理标准体系结构。在“数据安全—训练数据安全”部分，训练数据的质量治理与训练数据安全防护、安全标准等事项被并列为主要的治理内容。此外，工信部等八部门发布的《“人工智能 + 制造”专项行动实施意见》也明确提及开展“模数共振”行动，发布制造业高质量数据集建设指南，推动将基础数据转化为高质量行业数据集。

在整个人工智能技术治理体系中，数据质量保障是与数据安全保护同等重要的基础治理目标。不过，如若停留在政策文件层面的“高质量训练数据供给”等笼统表述，显然无法进一步推导出数据质量保障在数据投毒治理层面的具体表现形式。现行立法并没有对数据质量作出直接界定，主要还是通过各类技术标准细化数据质量保障的技术指标。在国家标准《网络安全技术　生成式人工智能预训练和优化训练数据安全规范》（GB/T 45652-2025）中，“通用安全要求”部分建议服务提供者对预训练、优化训练数据进行安全检测，修复或过滤被投毒数据。同时，该技术标准还针对预训练、优化训练数据处理活动等均提及“建立训练数据质量的评价机制”。全国数据标准化技术委员会在 2025 年发布的《高质量数据集质量评测规范》（TC609-5-2025-04）将“数据质量”界定为“在指定条件下使用时，数据的特性满足明确的和隐含的要求的程度”，并指出数据质量的评测指标应当是面向人工智能模型开发和训练的基本要求，具体的子指标包含格式规范性、安全规范性、标注规范性、结构完整性、内容真实性、内容一致性、类型一致性、内容干净性。此外，中国信息通信研究院联合另两家机构在该标准基础上发布了《人工智能高质量数据集建设指南》，

该指南将“高质量数据集”的核心特征解释为高价值应用、高知识密度、高技术含量，同时细化了人工智能数据集质量评估指标，具体包括完整性、规范性、准确性、及时性、一致性、稠密性、多样性、均衡性、相关性、原创性等。

如此看来，数据质量保障这一治理目标的实现似乎更应当放置于技术标准建构的话语体系中，而不是转向建构具体的数据质量法律规范。此种质疑不无道理，但也仅仅是观察到了数据质量保障的个别环节。事实上，一般场景下的数据质量保障依赖三类治理模式，即数据质量标准、数据质量管理工具以及数据质量管理流程。其一，数据质量标准化是为了明确数据处理者应当按照何种统一的技术标准收集、筛选、清洗、归并数据集合。在数据投毒领域，训练数据质量的标准化在一定程度上能够辅助数据处理者快速筛除不符合技术标准的训练数据，进而降低数据投毒风险。但是，数据质量标准并不是一成不变的，前述 TC609-5-2025-04 标准虽然与此前的国家标准《信息技术数据质量评价指标》（GB/T 36344-2018）采用相同的“数据质量”概念界定，但该技术标准发布于 2018 年，设置的数据质量评价指标以规范性、完整性、准确性、一致性、时效性、可访问性为限，客观上无法体现人工智能领域训练数据质量的技术实践发展变化。其二，数据质量管理工具是以“技术管理技术”为内核，强调通过特定的技术手段快速筛选和保障训练数据质量。当下，部分开源的数据质量探查工具能够辅助中小企业自动识别数据质量问题。[①]然而，不同企业的成本投入、技术水平并不相同，数据质量管理工具的选择也需要结合企业的商业实践需求进行选择。其三，数据质量管理流程则是为了衔接前两种治理方式，将整个数据质量管理活动转化为一整套完整业务规范流程，进而确保事前、事中以及事后均能够有效管控数据质量风险。不过，部分企业因忽视数据质量和业务规范能力有限等原因并未建构完善的数据质量管理流程，此时就需要法律介入，以义务规范的形式督促相关法律主体持续调整和优化数据质量管理流程。

① 参见Github官网，https://github.com/datachecks/dcs-core，访问日期：2026 年 2 月 23 日。

以数据质量保障义务为核心的数据投毒治理体系

数据质量保障在人工智能技术治理领域并非一个新兴概念，部分学者已经专门就数据质量保障制度提出不同的治理范式，大致可分为两类：一类研究方向侧重论证数据质量保障制度的基础理论，试图阐明数据质量保障所涉及的权利义务关系；[②]另一类研究方向则是回归数据质量保障所具有的技术特征，鉴于法律规范无法直接界定数据质量保障的具体要求，故而主张数据质量保障的制度方案应当是数据质量标准化。只不过在一些学者的论述中，数据质量保障与个人信息保护、隐私保护等基本权利相关联，数据质量保障存在“标准化与权利型”两种方式，法律规范的相关表述应当是以数据质量原则为基础，以实现数据正义为核心。[③]这两类研究方向并无不妥，但若从整体视角来看，不构成数据质量保障体系的全部内容。前一种研究方向主要是为了解决技术治理与法律治理的衔接问题，后一种研究方向则更加侧重数据质量保障的技术逻辑，借由数据标准化保障数据质量。

② 参见苏宇：《公共数据质量的制度保障》，《行政法学研究》2025 年第 4 期。

③ 参见高秦伟：《人工智能数据质量保障的规范性探究》，《学术研究》2025 年第 9 期。

现有研究遗漏了对不同数据质量保障方式之间逻辑关系的考察，进而导致相应的研究结论不自觉地偏向“专门立法论”或者“未来立法配套论”。而前文提及的“标准体系 + 技术工具 + 治理流程”的逻辑解决的是法律规则与技术规范之间如何衔接的问题，在面向数据投毒这类具体风险时，则需要进一步明确“法律规则究竟如何整合标准体系和技术工具这两类治理模式”。因此，本文提出的数据质量保障范式是为了在区分数据安全与数据质量的前提下明确数据质量保障的具

体制度建构。当然，这种范式更侧重对个体层面义务履行情况的判断和产业层面数据质量保障机制的衔接。在面向数据投毒治理时，该范式则需要结合训练数据质量风险的特殊性，转化为数据投毒治理的基本框架。

（一）基于数据质量保障范式的数据投毒治理框架

目前，有关数据质量保障制度的建构讨论大多以一般情形为预设前提，鲜有提及训练数据质量保障的特殊性。也有学者从数据错误等具体数据质量问题切入，提出对数据错误进行“条块化、纯净化与安全化等多维防护”，[①]并建立数据源头清理机制和数据应用阶段清理机制。但是，在数据投毒治理领域，这些研究结论缺乏对训练数据质量保障特殊性的回应，即忽略了对训练数据质量的标准考查。数据质量保障范式强调数据质量的判断标准并不是固定的，而是要面向训练数据的实际使用需求。在过去，数据质量的评价指标大多表述为完整性、一致性、真实性、准确性等，但在大模型预训练、优化训练、微调等场景下，这些评价指标并不能很好地反映训练数据的实践需求。同样地，在一般场景下，数据质量保障范式主要提供的是一种数据质量保障机制的建构框架和建构思路。在产业实践中，数据质量管理并不是某一技术环节的管理事项，而是需要结合主要业务内容设计相应的数据质量管理流程。并且，现行立法之所以没有专门制定数据质量保障制度，恰恰是因为不同场景、不同业务领域的数据质量保障需求并不完全一致。比如，《生态环境监测条例》第 25 条和第 42 条均提及建立健全监测数据质量管理制度，但因为这两条对应的义务主体分别是企事业单位和第三方技术服务机构，故而并没有在同一条款中予以规定。类似地，《政务数据共享条例》第 18 条提及政府部门应当建立健全政务数据全过程质量管理体系，第 38 条则强调政务数据共享主管部门应当加强对本行政区域内政务数据提供部门数据质量情况等的监督，这种全过程质量管理以及行政层级的内部监督本身亦是以政务数据处理需求为基础。

一方面，基于数据质量保障范式构建数据投毒治理框架时，应当以训练数据处理需求为导向。在一般场景下，数据质量的评价指标多表现为完整性、准确性、真实性等，但在人工智能治理领域，数据质量的评价依据是以个体的业务需求为导向。例如，在一般场景下，错误的文字信息显然不满足准确性、真实性等要求，而在以优化大模型识别错误文本能力的业务需求导向下，这类数据质量反而更高。并且，现行立法在提及数据质量管理相关内容时，多采用“建立健全数据质量管理制度”等笼统表述，并不提及具体如何建立健全这类制度。其根本原因便是，立法者不可能通过一条或者若干条法律条款囊括特定行业、特定领域所有的数据处理业务需求。即便采用禁止性条款建构相对具体的数据质量保障制度，督促义务主体充分履行数据质量保障义务，也会因禁止性规范过于笼统而与“建立健全数据质量管理制度”这类表述并无实质区别。进一步而言，训练数据质量评价标准的特殊性也决定了在判断义务主体是否履行数据质量保障义务时，不能直接以未采用数据质量标准作为单一标准，而是应当以义务主体是否结合自身需求设立和有效实施数据质量管理流程为判断依据。因为不同义务主体的数据质量管控能力并不完全一致，并且数据质量保障效果也不取决于数据质量管理流程的多寡，所以数据质量保障的重心是训练数据处理需求与数据质量管理流程的契合度。

另一方面，基于数据质量保障范式建构数据投毒治理框架还应当以数据投毒的防治能力为基础。在实践中，数据投毒的方式已经从最初直接向内部训练数据注入污染数据延伸至针对网络公共数据抓取投放污染数据。未来数据投毒方式究竟会如何变化尚不可知，但是数据投毒方式的多元化和技术更迭乃是必然。因此，数据质量保障所要求的数据质量管理流程是一种持续优化的内

① 彭中礼、左泽东：《数据错误的法律治理》，《四川大学学报》（哲学社会科学版）2024 年第 6 期。

部管理活动。这与数字法学领域常见的动态治理理念较为相似，但不同的是，数据质量保障范式不仅强调根据数据投毒方式的技术变化灵活调整数据质量管理流程，而且还强调数据质量管理流程要与实践中的数据投毒等数据质量风险相契合。之所以将数据质量与狭义的数据安全予以区分，是因为两者所要应对的风险有所不同：数据安全风险主要表现为数据泄露、数据损毁、数据篡改等，数据质量风险则主要表现为在特定场景下数据使用价值减损或者无法充分满足训练数据使用需求。这种风险特征实际上也影响到数据资源的商业价值，故而在建构数据投毒治理框架时，还需要将数据质量管理流程与数据资产化等外部制度予以衔接，避免数据质量管理成为纯粹的企业业务管理“负担”。

（二）个体层面的数据质量保障义务

在一般情形下，有些观点试图将数据质量与民事权利予以绑定，进而解释为何需要制定数据质量保障相关的法律制度，最常见的论据便是包含个人信息的数据资源一旦不准确，则可能对自然人的合法权益造成负面影响。这种思路实际上混淆了个人信息权益等民事权益保护与数据质量保障的保护目标，后者更多地表现为公法意义上的法定义务。特别是在数据投毒治理领域，数据质量影响的是人工智能产品和服务的功能，损害对象主要是研发者和服务提供者。当然，影响人工智能产品和服务功能确实会对自然人的个人信息或者其他合法权益产生影响，但这并非数据投毒治理的议题范畴，个人信息权益或者其他权益的减损问题完全可以通过违约责任或侵权责任的认定实现相应的治理目标。此外，数据投毒的损害结果虽然表现为私法层面人工智能产品和服务的功能减损，但数据投毒的行为主体难以有效识别，研发者、服务提供者能够采取的数据质量保障措施有限，事后追责的治理效果远不及事前预防。并且，从长远来看，数据投毒行为可能影响整个人工智能产业的创新进程，加之部分研发者、服务提供者因抢占市场竞争优势地位而忽视内部的数据质量保障管理，故而以研发者、服务提供者为义务主体设置公法意义上的数据质量保障义务更具有现实意义。该义务的法律依据便是《生成式人工智能服务管理暂行办法》第 7 条规定的“采取有效措施提高训练数据质量”，但由于数据质量的评价标准需要考虑义务主体的实际业务需求，所以设置数据质量保障义务的方式不是通过单独立法或者修法细化义务履行要求，而是结合数据质量标准、数据质量技术工具以及数据质量管理流程对义务履行标准进行“释法”。

基于数据质量保障范式所设置的义务履行标准是，义务主体根据自身训练数据使用需求规划、设计和有效实施相应的数据质量管理流程，并且该流程应当以现有的数据质量标准为指引，并根据义务主体的实际情况采取必要的数据质量管理技术工具。该义务的履行标准与数据安全保护义务履行常采用的“个案判断”“采取所有充足且必要的安全管理措施”等标准并不相同。数据质量保障并非同数据安全保障那般存在“安全”“危险”的二元评价指标，更强调数据是否能够满足业务实际需求。“个案判断”标准只会淡化设置数据质量保障义务的实践意义，“采取所有充足且必要的安全管理措施”标准则会对义务主体施加过重的义务负担，并且“充足且必要”本身的不确定性又会把义务履行标准引向“个案判断”。即便义务主体未采取常见的数据质量保障措施，但若其选择投保网络安全保险，借由保险人及其合作的网络安全服务机构的第三方服务保障数据质量，[①]同样可以视为其数据质量保障义务已经履行。从数据投毒治理的整体视角来看，义务主体履行数据质量保障义务仅是数据投毒治理措施的一部分，所以无法对义务主体施加过高的义务履行要求，而是需要通过义务主体的外在行为来判断其是否为了保障数据质量采取了相适配的管理措施。这里的“适配”主要包含三个要素：一是数据质量管理流程是否与主要业务活动同步

① 参见周瑞珏：《数据泄露风险治理中网络安全保险的介入路径》，《北方法学》2024 年第 2 期。

规划、同步设计、同步应用；二是在设计和应用数据质量管理流程之前，义务主体是否已经完成对数据质量保障需求的评估和认定；三是数据质量管理流程是否包含数据质量评估流程、数据质量检查流程、数据质量提升流程。

这种不以“采取充分且必要的管理措施”为内容的义务履行方式看似无法强制要求义务主体针对数据投毒风险采取相应管理措施，但数据投毒风险本身具有动态变化的特征，设置具体的强制性义务标准并无实际意义。更何况数据质量保障的根本目标是提升数据质量，只要义务主体采取了“相适配”的数据质量管理措施，就可以认定其已经履行了相应义务。为了预防数据投毒风险，常见的数据质量管理措施包括数据清洗、数据验证、后门验证、数据质量审计等措施。义务主体为了确保自身的业务发展等商业利益，在一定程度上也会自觉地采取这些措施确保数据质量，但这并不等同于这些管理措施都应当被视为数据质量保障义务的必要手段。因为单纯在数量上增加数据质量管理措施可能无法真正提升数据质量，反而会增加义务履行过程中的经济成本。此外，公法层面的数据质量保障义务是为了督促义务主体采取合理且适配的数据质量管理流程，而不是强制要求义务主体必须采取特定的数据质量管理措施。在实践中，数据质量保障也具有一定的专业技术门槛，在此种义务履行模式下，义务履行与否由义务主体举证更符合当下的数据投毒治理实践。

（三）产业层面的数据质量保障体系

数据投毒治理既依赖义务主体设置相适配的数据质量管理流程，还依赖产业层面的数据质量保障体系。数据投毒的发生通常是由外部行为人直接或间接导致的，义务主体内部的数据质量管理流程主要是为了预防数据投毒的个体事件，同时也是为了预防内部员工因故意或者疏忽导致训练数据被污染。但是，个体层面的数据质量保障义务并不足以有效预防数据投毒风险，结合数据投毒的作用方式和损害结果，还需要在产业层面形成与数据质量保障义务相衔接的数据质量保障体系。需要说明的是，在各个行业、各个场景下，数据质量保障的基本内涵、评价指标并不相同，并且数据投毒治理也与网络信息内容治理、关键信息基础设施安全保护、训练数据高质量供给保障等其他治理活动存在交叉，[①]所以数据质量保障体系的建构方式并非专门立法，而是借由释法或修法的方式将数据质量保障要求纳入人工智能技术治理的框架。

数据投毒方式包括直接对义务主体持有的训练数据进行篡改污染、利用已经投毒的开源训练数据或者开源大模型实施攻击活动、在网络上频繁发布特定的信息等，故而数据质量保障体系需要包含针对开源训练数据、网络信息内容治理等制度规范。具体而言，在网络信息内容治理领域，研发者、服务提供者应当依据《生成式人工智能服务管理暂行办法》第 9 条承担网络信息内容生产者责任，并针对数据投毒诱发的违法内容依照第 14 条规定采取停止生成、停止传输、消除等处置措施。同时，《网络信息内容生态治理规定》第 2 章专门列举了网络信息内容生产者的具体义务，研发者、服务提供者应当尽可能采取必要措施预防人工智能生成该规定第 6 条禁止的违法信息。此外，数据投毒的另一种作用方式是在公开的网络信息中增加特定网络信息的数量，进而让大模型在预训练、优化训练等环节将这些错误信息或者营销信息作为有效的训练数据，诱导大模型在用户输入相关指令时，生成投毒者导向的内容。但是，除网络谣言等错误信息外，营销类网络信息并不构成违法信息，在监管层面限制此类网络信息的内容的发布和传播缺乏正当性依据。所以，这类数据投毒治理仍然需要义务主体在其数据质量管理流程中清洗、筛选比例过多的营销类网络信息，并履行网络信息内容生产者义务和算法安全义务，避免用户因相信人工智能产品和服务输出的商户推荐信息、医疗机构推荐信息等而遭受人身财产损害。至于开源训练数据集可能

① 参见赵精武：《论人工智能训练数据高质量供给的制度建构》，《中国法律评论》2025 年第 1 期。

被数据投毒，鉴于开源活动本身属于一种技术民主实践，更适宜由开源社区管理者和参与者共同设立技术社区自治规范来预防数据投毒风险。

数据投毒的损害结果会因特定场景而有所不同，相应地，数据质量保障体系则需要在法律适用过程中增补有关数据质量保障的判断要素，包括但不限于关键信息基础设施安全保护、军事设施安全保护、自动驾驶行车安全监管规则等。在数据投毒治理中存在一种逻辑误区，即习惯性地将特定技术应用场景下人工智能系统风险治理机制与数据投毒治理机制混同。例如，在自动驾驶领域，自动驾驶系统因数据投毒而存在行车安全风险，此时将数据投毒治理机制解释为应当建构自动驾驶上路测试等监管制度。然而，即便没有数据投毒或者存在其他数据质量风险，也会得出同样的治理结论。因为自动驾驶系统本身就存在行车安全风险，诸如上路测试、系统安全保障等监管制度并不是专门为了数据投毒风险而设立，而是为了自动驾驶固有的技术风险而设立。所以，不是因为存在数据投毒风险而应当设立特定治理机制，而是需要将数据投毒风险作为治理机制适用的正当性依据。以关键信息基础设施运行安全为例，《网络安全法》等现行立法规定了诸如设置专门安全管理机构、组织推动网络安全防护能力建设等网络安全保护义务。但是，随着技术更新迭代，出现了诸如数据投毒这类风险，所以网络运营者的网络安全保护义务内容也需要延伸至数据质量保障领域。《关键信息基础设施安全保护条例》第 19 条提及运营者应当优先采购安全可信的网络产品和服务，那么这里的“安全可信”之内涵显然需要囊括人工智能系统能够具备一定的抵御数据投毒风险的能力。另外，该条例第 17 条要求运营者自行或者委托网络安全服务机构对关键信息基础设施每年至少进行一次网络安全检测和风险评估，数据投毒如若对人工智能系统安全造成威胁，那么相应的检测和评估事项应当包含数据投毒。由此可见，在产业层面，倘若仅因数据投毒风险而主张针对特定应用场景建构相应的治理机制，那么这种治理机制大概率会重复已有的法律规范。换言之，数据投毒的产业治理模式应当以法律解释为中心，细化现有网络安全、数据安全法律规范中对数据质量保障的基本要求。

结语

人工智能技术创新的发展过程必然伴随着诸多技术安全威胁，数据投毒风险便是其中之一。在数据投毒的治理过程中，应当关注法律规范与技术标准之间的话语体系差异，数据投毒风险的主要表现形式是训练数据质量的下降，进而减损大模型性能水平和干扰人工智能技术创新进程。在已有的数据要素治理体系中，数据安全、数据财产权等治理规则始终被广泛关注，但有关数据质量治理的讨论却屈指可数。数据质量保障并不单纯是技术治理层面的议题，同时也关系到人工智能技术治理范式的再优化。并且，从整体视角来看，影响人工智能创新的训练数据这一创新要素所面临的技术风险绝不仅限于数据投毒这类特殊技术威胁，还可能包括影响训练数据质量的其他风险。因此，在治理数据投毒的实践中，应当关注到数据投毒风险的技术原理、作用机制以及损害结果，推动研发者、服务提供者建构内部的数据质量管理流程，并与其他人工智能技术治理机制达成数据质量保障层面的内容衔接。同时，还应当立足于我国人工智能高质量训练数据供给的产业政策目标，基于数据质量保障范式，构建体系化、层次化的数据质量治理体系，以便更好地应对未来影响训练数据质量的不确定风险。

编辑　**孙冠豪**