财经法学 No. 3，2026
pp. 136 - 155

从数据删除到模型删除：人工智能监管转型的逻辑演化与中国路径

李汶龙*

内容提要：随着机器学习模型规模与复杂性不断扩张，数据训练的违法性已难以通过传统的数据删除或行为中止措施在规模上加以消解。在深度学习语境下，违法数据往往在训练过程中被压缩、嵌入并固化为模型参数与能力结构，使风险在数据处理行为停止后仍持续存在。围绕这一挑战，“模型删除”逐渐在美国、欧盟与韩国等法域发展为回应违法训练成果与模型残留风险的重要工具，其制度化可能在消费者保护法、数据保护法或个人信息保护法框架下实现，其功能形态则在权利延伸、惩罚性处置与纠正性监管之间呈现出差异化演进。在中国法语境下，单纯依赖删除权或行为控制型工具，难以回应模型能力固化所带来的持续风险。可通过对《个人信息保护法》第 61 条第 4 项的体系解释，将模型删除定位为以终止持续性违法状态为目标的纠正性监管机制。模型删除并非对既有数据保护逻辑的激进突破，而是其在模型时代得以自洽运作的制度补充。通过在现有法律框架内弥合行为违法与结构违法之间的断裂，中国有可能在保障创新与维护秩序之间建立灵活的模型治理路径。

关键词：模型删除　人工智能监管　算法没收　模型召回　机器去学习

一、问题的提出

近年来，以机器学习为核心的人工智能技术在计算结构和能力上实现了根本性突破，由此引

* 李汶龙，浙江大学光华法学院百人计划研究员、数字法治研究院研究员，英国爱丁堡大学数据、文化与社会研究中心研究员。

本文亦得益于 2025 年 10 月 2 日至 3 日在香港恒生大学传播学院举办的“中国个人资讯保护研讨会：网络私隐与数据道德”上所获得的宝贵意见与建设性反馈，在此致谢。

发的治理难题正在冲击现有法律制度的基础结构。〔1〕深度神经网络框架下，训练数据经过模型学习后不再以可识别条目或数据单元形式存在，而是转化为高度抽象的向量表征与参数结构。〔2〕数据从数据条目转化为模型能力的组成部分，使其进入一种既不可定位、不可拆分，又不可逆转的嵌入式存在状态。〔3〕这一技术机制深刻改变了数据治理所依赖的规范前提，使得以同意个人信息处理为始、删除个人信息为终的数据全生命周期风险控制路径面临系统性的功能弱化甚至失效。〔4〕

深度学习模型对训练数据进行复杂的非线性映射，导致个人信息以隐晦的形式嵌入模型参数中，难以直接定位和移除。〔5〕与传统数据库中明确的数据条目不同，模型内部参数并非按个体信息存储，而是散布在整个模型的权重和结构中，因此增加了删除的技术难度。再者，深度学习模型通常经历多轮迭代训练和增量学习，数据记忆不断融合与重塑，单纯删除某条数据对模型行为的影响难以预测，也难以保证彻底“遗忘”。〔6〕

传统数据保护制度以数据库范式为模型。〔7〕其逻辑基础在于，个人信息具有可识别存储载体，违法处理活动与其后果能够通过删除行为得以阻断，数据条目的消失即可宣告侵害结果终止。〔8〕在深度学习模型中，训练数据已不再具有这种条目形态，删除原始数据既难以影响模型能力，也无法消除模型在使用该数据过程中已形成的知识积累。被遗忘权等源自欧盟数据保护法的制度设计，在面对深度神经网络结构时，因缺乏对象识别性与执行可见性而逐渐陷入理论困境。〔9〕换言之，个人数据虽可在数据库层面删除，但模型内部的残留表征依然存续，由此形成的隐私风险排除了删除权作为完全救济手段的可能性。〔10〕当下计算科学领域，机器去学习和再学习成为显学，尝试通过技术优化的手段解决模型中个人信息残留难题。〔11〕但基

〔1〕 参见张凌寒：《深度合成治理的逻辑更新与体系迭代——ChatGPT 等生成型人工智能治理的中国路径》，载《法律科学（西北政法大学学报）》2023 年第 3 期，第 38 页。

〔2〕 See Lokke Moerel & Marjin Storm, *Do LLMs ‘Store’ Personal Data? This is Asking the Wrong Question*, IAPP (23 October 2024), https://iapp.org/news/a/do-llms-store-personal-data-this-is-asking-the-wrong-question, visited on 1 January 2026.

〔3〕 See Yeong Zee Kin, *Nature of Data in Pre-Trained Large Language Models*, Future of Privacy Forum (6 July 2025), https://fpf.org/blog/nature-of-data-in-pre-trained-large-language-models/, visited on 15 December 2025.

〔4〕 See Jennifer King & Caroline Meinhardt, *Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World*, Stanford University Human-Centered Artificial Intelligence (22 February 2024), https://hai.stanford.edu/policy/white-paper-rethinking-privacy-ai-era-policy-provocations-data-centric-world, visited on 6 January 2026.

〔5〕 See Antonio A. Ginart et al., *Making AI Forget You: Data Deletion in Machine Learning*, ACM Digital Library, https://dl.acm.org/doi/10.5555/3454287.3454603, visited on 6 January 2026.

〔6〕 参见林北征：《没有删除，只能遗忘：AI 大模型个人信息删除义务的解释与重构》，载《西安交通大学学报（社会科学版）》2024 年第 6 期，第 152 页。

〔7〕 See Hannah Ruschemeier, *Generative AI and Data Protection*, 1 Cambridge Forum on AI: Law and Governance 1 (2025).

〔8〕 See Jef Ausloos, *The Right to Erasure in EU Data Protection Law*, Oxford University Press, 2020, pp. 1 - 34.

〔9〕 See Eduard Fosch Villaronga, Peter Kieseberg & Tiffany Li, *Humans Forget, Machines Remember: Artificial Intelligence and the Right to Be Forgotten*, 34 Computer Law & Security Review 304 (2018).

〔10〕 See Michael Veale, Reuben Binns & Lilian Edwards, *Algorithms That Remember: Model Inversion Attacks and Data Protection Law*, Philosophical Transactions of the Royal Society A (15 October 2018), https://doi.org/10.1098/rsta.2018.0083, visited on 31 December 2025.

〔11〕 See Zachary Izzo et al., *Approximate Data Deletion from Machine Learning Models*, 130 Proceedings of Machine Learning Research 2008 (2021).

于现有技术，相关技术的开展和落实仍然存在挑战，并且实践中适用成本高，难以大规模开展。〔12〕

更为关键的是，训练数据的违法性正在从输入环节向模型成果层面延伸。〔13〕随着生成式人工智能依赖海量语料进行训练，模型训练数据往往涉及未经授权的著作权内容、来源不明的互联网文本、用户在不知情状态下提供的个人信息，以及缺乏合法处理基础的敏感数据。〔14〕即便违法训练材料被删除，模型能力仍将因既得训练效果而保留下来，构成一种事实上的持续违法状态。模型由违法数据训练而产生的能力增益不仅具有留存性，更具有显著经济价值，使违法输入优势直接转化为经济竞争优势。在这种结构下，违法训练行为若不触及模型本体，将无法实现法律救济效果。〔15〕

此外，模型在训练过程中形成的隐含表征可能再现训练文本片段、重构用户身份信息或表现特有表达风格，这些输出无法通过删除训练数据的方式加以限制。〔16〕这意味着，模型内部关于个人信息的记忆不以原始数据存续为前提，而以结构权重存续为条件，隐私侵害由此转化为模型产物风险，而非存储风险。传统删除机制并未触及这一层面，使删除数据即终结风险的制度逻辑失效。〔17〕

从更为宏观的法律框架层面，既有监管模式难以应对模型结构带来的违法后果。首先，在版权法上训练数据是否构成侵权利用、模型输出是否构成表达再现均已成为重要议题。在侵权成立的情形下，版权法亦可能通过停止侵害、赔偿损失或没收违法所得等方式对违法利用行为作出回应。〔18〕然而，在实际执法层面，版权没收规则并未直接延伸至对模型结构本身进行销毁或删除，目前主要还在聚焦训练阶段的数据合法性难题，以及 AI 生成物的版权侵权损害赔偿。其次，包括《个人信息保护法》在内的人格权法制度重心仍然聚焦于个人权益与信息处理行为。〔19〕无论是隐私权保护还是个人信息保护，其核心逻辑在于规范信息收集、使用与披露行为，以及保障数据主体的控制权。在这一框架下，违法处理行为可以被停止，相关数据可以被删除，但法律规范本身并未明确将模型能力视为独立的规制对象。再次，竞争法在模型层面亦存在一定讨论空间。例如，当企业通过违法数据获取训练优势并形成市场支配力时，相关行为可能被置于不正当竞争

---

〔12〕 See Xiaoyu Wu et al., *Unlearned but Not Forgotten: Data Extraction after Exact Unlearning in LLM*, Arxiv (2025), https://arxiv.org/abs/2505.24379, visited on 15 December 2025.

〔13〕 See Joshua A. Goland, *Algorithmic Disgorgement: Destruction of Artificial Intelligence Models as the FTCs Newest Enforcement Tool for Bad Data*, 29 Richmond Journal of Law & Technology 1 (2023).

〔14〕 See Mehtab Khan & Alex Hanna, *The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability*, 19 Ohio State Technology Law Journal 171 (2023).

〔15〕 See Jaydeep Borkar et al., *Privacy Ripple Effects from Adding or Removing Personal Information in Language Model Training*, Arxiv (2025), https://arxiv.org/abs/2502.15680, visited on 15 December 2025.

〔16〕 See Tiffany C. Li, *Algorithmic Destruction*, 75 SMU Law Review 419 (2022).

〔17〕 See Haley Higa, Suzan Bedikian & Lily Costa, *The Right to Be Forgotten Is Dead: Data Lives Forever in AI*, Tech Policy Press (20 May 2025), https://www.techpolicy.press/the-right-to-be-forgotten-is-dead-data-lives-forever-in-ai/, visited on 5 January 2026.

〔18〕 See Roy Baharad, *The Uneasy Case for Copyright Disgorgement*, Coase-Sandor Institute for Law & Economics Research Paper Series. 25 - 02 (23 June 2025), https://chicagounbound.uchicago.edu/law_and_economics/1046, visited on 3 March 2026.

〔19〕 参见赵精武：《生成式人工智能应用风险治理的理论误区与路径转向》，载《荆楚法学》2023 年第 3 期，第 47 页。

或滥用市场支配地位的分析框架之中。〔20〕然而，竞争法的核心在于市场结构与竞争秩序，其违法认定通常依赖对市场效果与竞争损害的证明。如何界定“模型产能”中因违法数据所形成的竞争优势，以及如何区分合法创新与不正当收益，仍存在高度争议。至于人工智能专门立法，其理论上最为直接面对模型层面的结构问题。欧盟《人工智能法》以及部分国家的专项规范，已开始将模型能力、风险等级与系统部署纳入独立规制对象。然而，全球范围内关于人工智能监管的路径呈现明显分化。

自2019年剑桥分析事件以来，〔21〕美国联邦贸易委员会（FTC）在多起执法案件中要求销毁基于违法数据训练的模型，旨在剥夺违法训练所产生的收益能力，防止违法成果固化为市场优势。在欧盟数据保护法框架下，执法机关通过纠正性措施要求模型下架或禁止使用，欧盟《人工智能法》更正式引入模型召回与撤回机制，使模型作为监管对象进入法律体系。〔22〕这些趋势表明模型删除并非监管构想，而是现实发生的新型结构性应对措施，其功能定位超越传统隐私保护，转向违法成果遏制、风险治理与市场纠偏。

基于上述困境，本文提出重新理解“删除”这一概念，其内涵不应囿于个人信息层面的消除，而应上升为对模型本体风险的解除。模型删除机制并非仅对删除权的技术延伸，而是一种面向模型成果的结构性治理手段。其功能不在于恢复个体信息控制，而在于遏制违法收益、修复竞争秩序、实现风险清除，并保障数据处理体系与市场结构的长期稳定。面对深度学习模型的违法成果、隐私残留性与能力永续化风险，传统治理路径已无法满足体系需求，从数据条目控制走向模型产物控制。基于此，本文将围绕模型删除的内涵边界、类型划分与制度属性展开，在梳理美国、欧盟等制度路径的基础上，提出模型删除在中国法律体系中落地的解释空间与规范路径。

二、模型删除的理论基础

（一）模型删除的定义

模型删除是指对人工智能模型本体采取移除、擦除或结构性修改措施，目的在于消除模型中因训练数据违法、处理基础缺失或个人信息残留而产生的持续性风险，弥补传统以删除原始数据为中心的数据治理工具在深度学习语境下的功能不足。与仅针对数据存储或处理行为的删除不同，模型删除直指作为技术产物及风险载体的模型，具体形式既包括对模型的彻底销毁，也包括将其从市场中撤回、停止部署，或通过再训练、去学习等技术对模

〔20〕 See Vaibhav Srikaran, *Beyond Data Deletion: Addressing Anticompetitive Conduct in the Era of Machine Learning*, 20 Washington Journal of Law, Technology & Arts 139 (2025).

〔21〕 See Federal Trade Commission, *FTC Issues Opinion and Order Against Cambridge Analytica For Deceiving Consumers About the Collection of Facebook Data, Compliance with EU-U.S. Privacy Shield*, (6 December 2019), https://www.ftc.gov/news-events/news/press-releases/2019/12/ftc-issues-opinion-order-against-cambridge-analytica-deceiving-consumers-about-collection-facebook, visited on 6 December 2019.

〔22〕 See Alessio Tartaro, *When Things Go Wrong: The Recall of AI Systems as a Last Resort for Ethical and Lawful AI*, 5 AI and Ethics 253 (2025).

型结构加以调整。

现有文献与监管实践已出现多种相近但侧重点不同的概念安排。美国法中常使用“算法没收”的表述，强调通过强制删除或销毁基于非法数据训练形成的模型剥夺违法行为所带来的技术能力与经济利益，其制度逻辑主要植根于消费者保护与不公平竞争规制。〔23〕欧盟《人工智能法》则采用“模型撤回”与“模型召回”的术语，侧重于（产品）安全、健康或基本权利风险，其关注重点在于模型的市场可用性，而非模型内部参数结构的技术性处理。欧盟数据保护委员会（EDPB）亦在数据保护法相关指引中提出“模型删除”与“模型匿名”等概念，用以回应个人数据深度嵌入模型所带来的合规挑战。〔24〕与此同时，计算科学领域近年来围绕模型再训练与机器去学习展开大量讨论，试图通过技术方式削弱或消除特定数据对模型行为的影响，以在不完全摧毁模型的前提下实现风险缓释。〔25〕本文采用“模型删除”作为统摄性概念，利用其概念弹性将不同法域和学科中围绕“模型层面风险消除”所发展出的制度与技术路径纳入同一分析框架，作为后文比较不同治理模式与制度逻辑的统一解释基础。

本文将模型删除概括为三种主要形式：物理损毁式、市场撤回式与技术重配式。物理损毁式模型删除是指通过彻底销毁模型本身及其所包含的参数、权重与衍生成果，使模型在技术与法律意义上均处于不可恢复、不可再利用的状态。这一形式具有高度的不可逆性，能够从根本上切断模型中潜在的个人信息残留或违法训练成果。其功能类似于知识产权法中对侵权假冒商品的销毁，或海关监管中对非法出版物、不合格进出口货物及走私物品的无害化处置。〔26〕目前，此类模型删除措施主要出现在美国，FTC 在多起案件中要求企业在删除违法获取的训练数据之外，对相关模型及产品整体予以销毁。市场撤回式模型删除是指在监管机关或法院要求下，将存在安全风险、合规缺陷或基本权利隐患的模型从市场中撤回，禁止其继续提供服务或进行商业化部署。模型在技术上未必被立即销毁，但其市场流通与实际使用被中止，风险得以通过“不可用”状态加以控制。该模式与产品责任法中的缺陷产品召回机制高度相似，强调快速消除对公众安全和基本权利的威胁，具有实施成本相对较低、见效迅速的特点。〔27〕欧盟《人工智能法》第 97 条所确立的模型撤回与召回机制即属此类，但尚未出现成熟执法实践。技术重配式模型删除则是通过算法与工程手段对既有模型进行再训练、微调或机器去学习，以削弱或消除特定违法或敏感数据对模型行为的影响，同时尽可能保留模型的整体功能与性能。〔28〕这一路径在理论上有助于在隐私保护与创新激励之间取得平衡，但目前技术实现难度较高且成本不容

---

〔23〕 See Margot Kaminski, *AI Disgorgement or AI Recalls: A Trip Down Remedy Lane*, Jowell (2025), https://scholar.law.colorado.edu/cgi/viewcontent.cgi?article=2755&context=faculty-articles, visited on 15 December 2025.

〔24〕 See European Data Protection Board, *EDPB Opinion on AI Models: GDPR Principles Support Responsible AI*, (18 December 2024), https://www.edpb.europa.eu/news/news/2024/edpb-opinion-ai-models-gdpr-principles-support-responsible-ai_en, visited on 15 December 2025.

〔25〕 See Sijia Liu et al., *Rethinking Machine Unlearning for Large Language Models*, 7 Nature Machine Intelligence 181 (2025).

〔26〕 参见赵精武：《从保密到安全：数据销毁义务的理论逻辑与制度建构》，载《交大法学》2022 年第 2 期，第 28 页。

〔27〕 参见王利明：《关于完善我国缺陷产品召回制度的若干问题》，载《法学家》2008 年第 2 期，第 69 页。

〔28〕 See Bjørn Aslak Juliussen, Jon Petter Rui & Dag Johansen, *Algorithms that Forget: Machine Unlearning and the Right to Erasure*, 51 Computer Law & Security Review 105885 (2023).

忽视。机器去学习更多停留在科研探索与工程优化层面，目前并未发展出成熟、可复制的合规解决方案。

（二）模型删除的形式及法律基础

模型删除并非一项无争议的治理工具，在不同法域中存在显著分歧。〔29〕一方面，传统数据保护法所明确确立的仅是数据主体请求删除其个人数据的权利，该权利是否、以及在何种程度上可以延展至模型层面尚缺乏共识。另一方面，从财产法与宪法保障的角度看，人工智能模型及其参数结构、算力投入通常被视为开发者或部署者的重要财产性利益，尤其是在要求对模型进行物理销毁的情形下，其法律效果已接近国家以强制力剥夺私有财产，因此需接受严格的正当性与比例性审查。再者，有学者指出，由于训练数据的质量、多样性、合法性在很大程度上决定了模型的表现、公平性以及幻觉等本质问题，AI 监管的核心仍是数据治理。〔30〕模型删除是上述治理思路的平行路径，直接在模型层面寻找治理和救济进路。下文将结合比较法与现有实践，梳理模型删除在不同法理脉络中的制度基础，以揭示其并非单一权利或制裁，而是多种规范逻辑在人工智能语境下的交汇结果。

1. 模型没收：违法收益剥夺的延伸路径

美国法中，模型删除最具代表性的制度形态是“算法没收”。其法律基础主要源自 FTC 在消费者保护领域的执法权力。《联邦贸易委员会法》（FTC Act）第 5 条授权 FTC 针对不公平或欺骗性商业行为采取“合理针对违法行为的救济措施”。从法律性质上看，没收违法所得属于行政法或刑法意义上的强制性剥夺措施，其目的在于消除违法行为所带来的经济激励，防止违法者通过不正当手段持续获利。〔31〕这一逻辑不同于民法上的不当得利返还，后者强调受损方的利益恢复，而非对违法者的制裁与威慑。

自 2019 年剑桥分析案以来，FTC 逐渐将这一授权解释为包括强制删除基于违法获取数据训练形成的算法或人工智能模型。尤其是在 2021 年美国最高法院于 AMG Capital Management 一案中明确否定 FTC 依据第 13（b）条追缴金钱性违法所得的权力之后，〔32〕FTC 在执法实践中转而更多依赖非货币性的结构性救济，其中模型删除成为最具代表性的替代工具。

需要指出的是，现有 FTC 执法文件中对模型没收的法律基础与比例性论证并不充分。〔33〕有学者指出，算法没收的兴起，在相当程度上是一种制度“变通”，即在金钱性制裁受限的背景下，通过摧毁违法训练所形成的技术能力实现对违法收益的实质性剥夺。这也决定了模型没收在美国

〔29〕 See Jeremy Straub, *Algorithmic Disgorgement is Bad for Science and Society*, Lawfare (12 June 2023), https://www.lawfaremedia.org/article/algorithmic-disgorgement-is-bad-for-science-and-society, visited on 15 December 2025.

〔30〕 See Julie E. Cohen, *Public Utility for What?: Governing AI Datastructures*, 28 Yale Journal of Law & Technology 135 (2025).

〔31〕 See Benjamin Raue, *Disgorgement of Profits: Distributive and Deterrant Logics*, in Franz Hofmann & Franziska Kurz eds., Law of Remedies: A European Perspective, Intersentia, 2019, pp. 153 - 167.

〔32〕 See AMG Capital Management, LLC, et al. *v.* Federal Trade Commission, 593 U.S. ____ (2021), 141 S.Ct. 1341 (2021).

〔33〕 See Lydia Belkadi & Catherine Jasserand, *From Algorithmic Destruction to Algorithmic Imprint: Generative AI and Privacy Risks Linked to Potential Traces of Personal Data in Trained Models*, https://blog.genlaw.org/CameraReady/5.pdf, visited on 15 December 2025.

法中更接近一种惩罚性与威慑性并重的行政措施，而非权利救济或合规修复手段。

2. 模型“召回”：产品安全与风险控制逻辑

与美国不同，欧盟发展出另一条规范进路，即“模型召回”或“模型撤回”。其中，召回通常针对已部署系统的回收或停用，而撤回则防止尚未投放市场的系统进入流通，二者共同目标在于通过限制模型的可用性来实现公共安全与基本权利的保护。这一模式下，模型删除的核心目的并非惩罚违法行为本身，而是尽快阻断风险。模型召回与撤回的法律基础主要源自产品安全法与消费者保护法，并通过行业监管规则扩展适用于人工智能产品。〔34〕传统上，产品召回是一种以风险预防为导向的制度安排，当产品存在安全缺陷、功能失效或违反强制性规范时，监管机构可以要求生产者对产品进行召回、修复或销毁。〔35〕

随着人工智能系统逐步嵌入医疗、交通、金融等高风险领域，包含模型的软件产品亦被纳入类似的风险治理框架。例如，医疗器械监管中，若算法或软件存在设计缺陷，足以影响诊断或治疗安全，即需通过召回或修正措施予以纠正。〔36〕欧盟《人工智能法》在此基础上进一步制度化了模型召回与撤回机制。按照该法第 79 条的设计，当人工智能系统，尤其是高风险系统，被发现对健康、安全或基本权利构成重大风险时，提供者负有采取撤回或召回措施的义务，包括停止部署、终止服务或禁止使用，并向监管机构报告相关情况。〔37〕

3. 模型禁止/中止处理个人数据：数据保护法的行为控制路径

实际执法监管中，模型删除更多通过“禁止或限制处理”的方式间接实现。当人工智能模型涉及非法、未经同意或缺乏合法基础的个人数据处理时，监管机关可以依据其纠正性权力要求停止相关数据的使用。〔38〕实践中这一措施往往表现为对涉案模型的“下架”或“停止运行”，从而防止模型继续利用违法数据。以欧盟《通用数据保护条例》（GDPR）为例，第 58（2）（d）条授权监管机关命令控制者或处理者以指定方式使数据处理活动合规，第 58（2）（f）条则明确允许采取临时或永久性的限制措施，包括禁止处理。这些权力通常被解释为可以要求暂停相关服务或从市场中移除模型。例如，OpenAI 在 2023 年进入欧洲市场时并未有效开展 GDPR 合规，意大利数据保护监管机构于 2023 年针对 ChatGPT 发布的临时禁令即体现了这一逻辑。需要注意的是，这一路径的规范重心仍然停留在“处理活动”的控制层面，其法律效果主要是使模型在特定场景中不可用，而非对模型本体进行结构性处置。

4. 模型擦除：纠正性权力的结构性延伸

除上述路径外，部分法域中还出现了另一种更具争议性的模型删除逻辑，即在不以训练数据

〔34〕 参见杨慧：《论缺陷产品召回制度对消费者权益的保护》，载《安徽大学学报（哲学社会科学版）》2007 年第 4 期，第 88 页。

〔35〕 参见李友根：《论产品召回制度的法律责任属性——兼论预防性法律责任的生成》，载《中国检察官》2012 年第 5 期，第 78 页。

〔36〕 See Branden Lee et al., *Early Recalls and Clinical Validation Gaps in Artificial Intelligence-Enabled Medical Devices*, 6 JAMA Health Forum e253172 (2025).

〔37〕 See Alessio Tartaro, *When Things Go Wrong: The Recall of AI Systems as a Last Resort for Ethical and Lawful AI*, 5 AI and Ethics 253 (2025).

〔38〕 See Paweł Hajduk, *The Powers of the Supervisory Body in the GDPR as a Basis for Shaping the Practices of Personal Data Processing*, 45 Review of European and Comparative Law 57 (2021).

违法性为前提的情况下，直接要求对模型进行擦除或销毁。欧盟数据保护委员会（EDPB）在其关于人工智能模型的指引中明确提及，监管机关除数据删除外，还可以基于其纠正权要求实施“模型擦除”。该主张并未明确其具体法律机理，且 GDPR 第 58 条赋予监管机关的权力本身具有高度概括性，使得该路径目前仅存在于理论层面。

类似趋势亦可在美国与韩国的执法实践中观察到。自 2024 年 Rite Aid 案以来，FTC 在部分案件中开始要求删除并非基于违法训练数据的模型，而是将模型删除作为治理模型滥用与系统性风险的工具。在 Kakao Pay 案中，因韩国公民个人数据传输至中国，韩国个人信息保护委员会于 2025 年针对 Kakao Pay、苹果与阿里巴巴作出的处罚决定中，直接要求对涉案算法进行删除，但并未对其法律基础作出充分说明，仅以“充分解决违规事由”为由加以概括。〔39〕这一做法更接近于将模型视为需被清除的风险载体或“违禁品”，而非通过下架或再训练加以修复。

5. 模型去学习：技术性纠正与法律义务的耦合

机器去学习旨在通过算法手段削弱或移除特定数据对模型行为的影响，使模型状态尽可能接近“未曾学习该数据”的情形。〔40〕相较于删除源数据并完全重新训练模型的高昂成本，去学习被视为一种更具合理性的技术替代方案。从功能上看，其目标在于消除模型中的数据影响痕迹，亦可被视为模型删除法律机制的技术实现路径。

规范层面，模型去学习可以与多种法律义务形成耦合。一方面，个人信息保护法中的更正权与删除权虽在模型语境下面临适用困难，但当模型因使用违法数据或撤回同意的数据而持续输出不准确或侵权性结果时，数据控制者原则上负有通过技术手段消除影响的义务。〔41〕另一方面，从民法角度看，模型去学习亦可被理解为瑕疵补正义务在算法与数据场景中的延伸。当产品或服务未能达到合同所期的性能或安全标准时，提供者负有修复或替代义务。EDPB 在 2024 年的相关指引中亦认可，通过技术上可行的模型调整方式实现合规删除的可能性，尽管其并未对“删除”所对应的具体技术路径作出限定。〔42〕

三、国外制度发展与比较视野

（一）美国算法没收制度

自 2019 年起，美国逐渐形成了一套有别于传统隐私法的新型救济机制——算法没收。这一

〔39〕 서인숙, 개인정보위, 개인정보 무단 국외 이전한 카카오페이 · 애플에 총 83억 7,520만 원 과징금 · 과태료 부과, https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=10955，2025 年 11 月 17 日访问。

〔40〕 See Haibo Zhang et al., *A Review on Machine Unlearning*, Arxiv (2024), https://arxiv.org/abs/2411.11315, visited on 11 October 2025.

〔41〕 See Taner Kuru, *Lawfulness of the Mass Processing of Publicly Accessible Online Data to Train Large Language Models*, 14 International Data Privacy Law 326 (2024).

〔42〕 See European Data Protection Board, *EDPB Opinion on AI Models: GDPR Principles Support Responsible AI*, (18 December 2024), https://www.edpb.europa.eu/news/news/2024/edpb-opinion-ai-models-gdpr-principles-support-responsible-ai_en, visited on 11 October 2025.

概念存在歧义，有学者指出，实际上基于 FTC 执法形成的算法救济机制本质上并非是返还或剥夺不当得利，〔43〕因为传统的返还违法所得衡平性救济要求因果限度（被返还的财产或利润必须与违法行为具有明确因果关系）、比例要求（返还的范围和程度与实际的不当得利成比例）以及收益关联等标准，而美国 FTC 实施的模型删除命令更接近一种惩罚性、预防性的行政销毁手段。美国学者将 FTC 执法称为“无劣字节规则”，即只要模型的训练数据包含任何部分非法数据，就必须彻底删除整个模型，因此脱离了返还或剥夺不当得利理论的范畴。〔44〕

在 FTC 算法没收的发展历程中，2021 年美国最高法院对 AMG Capital Management 一案的判决，对该救济形式的存续与定位具有重大影响。在该案中，原告通过隐藏条款发放高利率短期贷款，被 FTC 依据《联邦贸易委员会法》第 13（b）条起诉并要求返还约 12.7 亿美元。但是，美国联邦最高法院一致裁定该条款仅授权发布禁令，不授权法院判令金钱性救济，例如返还或追缴利润。在 AMG 案之前，FTC 依据《联邦贸易委员会法》第 13（b）条，主张其有权要求被告返还不法所得。然而，美国最高法院在判决中的结论直接导致传统意义上的货币性利润剥夺失去了法定基础。〔45〕该判决一度使 FTC 陷入执法困境，其赖以震慑违法行为的经济制裁权被削弱。算法没收正是在这一背景下 FTC 为弥补货币性没收权被撤销而创造出来的替代工具。FTC 巧妙地以“删除模型”替代“追缴利润”，一方面规避了“事后金钱救济”（性质上属于防止持续或再犯危害，而非追溯既得利润），另一方面重新定义了“剥夺”的对象，从获利转化为算法和技术成果。

总体来说，美国的算法没收经历了三个阶段：最初以隐私执法为焦点，继而扩展到儿童保护与公平使用，最终进入广义的 AI 治理框架，不再限于违法获得训练数据及其影响（即模型）的修正。

算法没收最初出现在剑桥分析案。在该案中，剑桥分析公司通过 Facebook 平台的第三方应用程序收集了超过五千万名用户及其好友的数据，并在用户不知情的情况下建立心理与政治倾向评估模型，用于政治广告定向和选民画像，构成了《联邦贸易委员会法》第 5 条下的欺骗性行为。〔46〕FTC 在最终命令中要求删除所有非法收集的数据，以及任何由这些数据“直接或间接派生出的工作成果”，包括“任何算法或方程式”。这一命令奠定了算法没收的制度雏形，标志执法目标从数据层面扩展至算法产物层面，不仅要删除非法数据本身，还要摧毁利用这些数据训练出的模型，防止违法者从中获得任何技术或经济利益。2021 年，FTC 在 Everalbum 案中再次援引

〔43〕 See Daniel Wilf-Townsend, *The Deletion Remedy*, 103 North Carolina Law Review 1809 (2025). 值得注意的是，美国 FTC 法下讨论的返还或剥夺不当得利原则，需与大陆法系民法中的“不当得利返还”进行区分。前者是以公共秩序为导向的行政性、制裁性剥夺，而后者是以权利平衡为导向的私法性、补偿性返还。虽然在中文语境下译为“不当得利”，逻辑上更接近于行政法或刑法上的没收，其本质不是为了赔偿，而是削弱违法行为的后果，并断绝其衍生效应。换言之，其内核是要求企业交出因违法数据处理构建的“违法果实”，不论受害个体是否能获赔，这与民法中返还利益、实现个体之间利益平衡的思路完全不同。

〔44〕 See Daniel Wilf-Townsend, *The Deletion Remedy*, 103 North Carolina Law Review 1809 (2025).

〔45〕 See AMG Capital Management, LLC, et al. *v*. Federal Trade Commission, 593 U. S. ____ (2021), 141 S. Ct. 1341 (2021).

〔46〕 See European Data Protection Board, *EDPB Opinion on AI Models: GDPR Principles Support Responsible AI*, (18 December 2024), https://www.edpb.europa.eu/news/news/2024/edpb-opinion-ai-models-gdpr-principles-support-responsible-ai_en, visited on 15 December 2025.

相同原则，但扩展了适用范围。〔47〕Everalbum 是一款云端照片储存与整理应用，其人脸识别功能默认开启，并将用户上传的照片及面部标识数据用于自有算法模型训练，且未取得用户明确同意，违反了隐私透明与选择权义务。〔48〕FTC 认定，这一行为构成欺骗性陈述与不公平数据处理，并在和解命令中要求 Everalbum 销毁所有未经同意收集的生物识别数据，以及“任何全部或部分基于此类数据开发的模型或算法”。〔49〕时任专员罗希特·乔普拉（Rohit Chopra）将此举描述为“欺诈之果必须被没收”，彰显其惩罚性与威慑功能。〔50〕2022 年，FTC 在 Weight Watchers 案中将该机制用于儿童隐私违规领域。〔51〕随后的 Edmodo 案中，FTC 再次要求销毁未获父母同意收集的儿童数据及基于这些数据的算法。〔52〕

Rite Aid 案被视为美国算法没收的分水岭。该公司因在药店部署人脸识别系统时存在系统性歧视与安全漏洞，被认定违反《联邦贸易委员会法》第 5 条规定的禁止不公平行为。〔53〕尽管该案中并无非法数据收集行为，FTC 仍命令销毁数据与模型。这标志着算法没收从基于数据违法的“数据型”没收，扩展至针对“使用不当”的“使用型”没收：即便模型合法训练，如果其使用方式被认为不公平或有害，仍可成为摧毁对象。

美国司法层面，算法没收仍主要停留于行政与和解阶段，其跨领域示范效应已经显现。2022 年得克萨斯州诉 Meta 案中，Meta 被指控多年间一直使用面部识别技术处理上传到 Facebook 的照片，Meta 的“标签建议”功能未经适当同意，采集、分析和存储了得克萨斯州居民的面部几何数据，得克萨斯州州检察长请求法院命令销毁所有基于非法生物识别数据训练的算法。最终该案未采纳该措施，但其提出本身标志着州执法者试图将算法没收司法化。〔54〕

（二）欧盟模型删除与召回/撤回

欧盟的“模型删除”机制经历了一个渐进性的立法与监管演化过程。模型删除在欧盟并非

〔47〕 See Federal Trade Commission, *FTC Finalizes Settlement with Photo App Developer Related to Misuse of Facial Recognition Technology*, (7 May 2021), https://www.ftc.gov/news-events/news/press-releases/2021/05/ftc-finalizes-settlement-photo-app-developer-related-misuse-facial-recognition-technology, visited on 7 May 2021.

〔48〕 Ibid.

〔49〕 Ibid.

〔50〕 See Rohit Chopra, *In the Matter of Everalbum and Paravision*, (8 January 2021), https://www.ftc.gov/system/files/documents/public_statements/1585858/updated_final_chopra_statement_on_everalbum_for_circulation.pdf, visited on 10 October 2025.

〔51〕 See Federal Trade Commission, *FTC Takes Action Against Company Formerly Known as Weight Watchers for Illegally Collecting Kids' Sensitive Health Data*, (4 March 2022), https://www.ftc.gov/news-events/news/press-releases/2022/03/ftc-takes-action-against-company-formerly-known-weight-watchers-illegally-collecting-kids-sensitive, visited on 4 March 2022.

〔52〕 See Federal Trade Commission, FTC Says Ed Tech Provider Edmodo Unlawfully Used Children's Personal Information for Advertising and Outsourced Compliance to School Districts, (22 May 2023), https://www.ftc.gov/news-events/news/press-releases/2023/05/ftc-says-ed-tech-provider-edmodo-unlawfully-used-childrens-personal-information-advertising, visited on 22 May 2023.

〔53〕 See Federal Trade Commission, *Rite Aid Banned from Using AI Facial Recognition After FTC Says Retailer Deployed Technology without Reasonable Safeguards*, (19 December 2023), https://www.ftc.gov/news-events/news/press-releases/2023/12/rite-aid-banned-using-ai-facial-recognition-after-ftc-says-retailer-deployed-technology-without, visited on 19 December 2023.

〔54〕 See Zach Despart, *Texas AG Ken Paxton Says Google Will Pay Texas $1.4 Billion to Settle Privacy Suit*, https://abc13.com/post/texas-attorney-general-ken-paxton-says-google-will-pay-14-billion-settle-privacy-suit/16448859/, visited on April 30 2026.

单一制度创新，而是通过 GDPR 与《人工智能法》两大框架的交叉演化，形成了一个兼具权利保障与市场安全监管的双重治理体系。这一框架与美国 FTC 使用的算法没收有显著区别：后者属于制裁性剥夺，前者的形式更为多元，法律机理也更复杂。〔55〕

在 GDPR 框架下，模型删除有两种主要路径。一种是基于原则导向的扩张解释：通过对数据最小化、存储限制、合法性等原则进行系统阐释，认为当个人数据深度嵌入模型而持续产生违法影响时，删除模型即为避免“非法处理”的合理延伸。另一种则是权力导向的实践性解释：数据保护机关依据第 58 条的“命令删除”权力，直接适用其于模型层级。例如，意大利数据保护监管机关针对 ChatGPT、Replika〔56〕以及 DeepSeek〔57〕多家 AI 模型厂商开展“暂时封禁”的措施。三起案件中，监管机关采取临时禁令的法律逻辑具有高度一致性，均围绕 GDPR 关键条款的合规展开。以针对 OpenAI 的案件为例，其禁令主要基于涉嫌违反 GDPR 若干关键条款，包括缺乏合法的数据处理依据以及未履行充分透明度义务等。随后出现的 Replika 案件则将监管重点转向未成年人数据保护问题，认为其未建立充分的年龄验证与防护机制。在 DeepSeek 案件中，争议焦点则集中于个人数据存储于中国，涉及数据跨境传输的合规要求。此外，德国柏林数据保护专员亦在数据跨境案件中针对国产模型 DeepSeek 提出“访问屏蔽”要求，但并非基于 GDPR 第 58 条，而是以欧盟《数字服务法》（Digital Services Act，DSA）对数字平台提出的非法内容治理义务为法律基础。〔58〕

执法层面，模型删除已被各国监管机关以 GDPR 执法的形式探索性引入。目前欧盟执法机构日益将“模型移除”作为一种以行为矫正为主的软性删除机制，体现欧盟更为注重合规修复的监管理念。但是，由 EDPB 基于第 58 条修正权力提出的模型删除理念仍有可能在未来衍生出不同形式的模型删除，包括美国惩罚性摧毁算法收益的做法。

欧盟《人工智能法》从另一维度确立了模型删除的制度化基础。其第 79 条引入“撤回”“召回”概念，将 AI 系统置于类似产品安全与市场监管的架构下。当 AI 系统被判定对“健康、安全或基本权利”构成风险时，市场监督机构可命令其停止服务、撤出市场或销毁。模型删除不再仅因数据违法而触发，而可基于风险性或危害性本身。即便数据来源合法，只要模型风险无法改正，也可被命令撤回或召回。这种逻辑与 GDPR 下要求模型暂时封禁或下架的做法相似，但触发条件更宽，体现了从数据权利侵害到公共风险治理导向的延伸。《人工智能法》仍处于生效初期，执法面临外部地缘政治压力（主要来自美国）以及内部整改呼声（欧盟委员会提出的 Digital

〔55〕 See Bjørn Aslak Juliussen, Jon Petter Rui & Dag Johansen, *Algorithms that Forget: Machine Unlearning and the Right to Erasure*, 51 Computer Law & Security Review 105885 (2023).

〔56〕 See Eleonora Curreli & Laura Liguori, *The Italian Data Protection Authority Blocks AI Chatbot Replika Due to Endangerment of Minors and Vulnerable People*, https://www.mondaq.com/italy/privacy-protection/1290994/the-italian-data-protection-authority-blocks-ai-chatbot-replika-due-to-endangerment-of-minors-and-vulnerable-people, visited on 30 April 2026.

〔57〕 See Pierluigi Paganini, *Italy's Data Protection Authority Garante Blocked the DeepSeek AI Platform*, https://securityaffairs.com/173680/security/italys-data-protection-authority-garante-blocked-deepseek.html, visited on 30 April 2026.

〔58〕 서인숙, 개인정보위, 개인정보 무단 국외 이전한 카카오페이 · 에플에 총 83억 7,520만 원 과징금 · 과태료 부과, https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=10955, 2025 年 11 月 17 日访问。

Omnibus 简化案涉及该法〔59〕），因此暂未出现模型召回或撤回的案例。

（三）韩国个人信息保护执法

模型删除在欧盟和美国之外尚未形成明显的扩张趋势。不过，韩国个人信息保护委员会（PIPC）针对苹果、Kakao Pay 与阿里巴巴作出的决议具有代表性，标志着“模型删除”向东亚地区的延展。

韩国关注模型层面的治理可以追溯至 2021 年“이루다（Iruda）”事件，PIPC 首次依据韩国《个人信息保护法》（PIPA）对 AI 聊天机器人开发商 ScatterLab 实施制裁，认定其未经同意将 KakaoTalk 聊天记录用于训练 AI 模型，违反数据最小化与合法处理原则。这是韩国首次适用 PIPA 于 AI 系统，在韩国隐私法律史中具有里程碑意义。〔60〕但是，PIPC 的执法仅限于删除非法收集的个人信息与停止违规处理活动。尽管行政机关后续解释中提及“AI 系统应停止运行”，强调不应再利用违法数据训练模型，但官方文书并未出现“模型删除”或“销毁算法”的直接表述。即 PIPC 要求 ScatterLab“删除原始个人数据并防止进一步使用”，并未明文命令其摧毁或擦除模型本身。〔61〕严格意义上说，“Iruda 案”启动了对算法层面数据残留的监管意识。

明确提出并执行“模型删除”指令的是 PIPC 在 2025 年针对 Kakao Pay、Apple 与 Alipay 作出的裁决。〔62〕PIPC 认定 Alipay 在未获用户同意的情况下，利用从 Kakao Pay 传输的用户数据构建并运行了“缺乏资金风险评分模型”，据此进行持续性信用评估。这一案件的特别之处是阿里巴巴被要求删除涉案算法模型，而非仅暂停服务。PIPC 指出该措施目的在于“充分解决违规事宜”。这是迄今为止韩国个人信息保护制度中首次出现“销毁”AI 模型本身的正式命令。由此，2025 年 Kakao Pay 案可视为韩国在实质意义上首次实施法律意义上的“模型删除”。

这一发展不是对欧盟规则的被动移植，而是布鲁塞尔效应下的一种“自觉趋同”。〔63〕韩国的模型删除实践，一方面是对欧盟监管经验的主动吸纳，另一方面也是面对本国数字伦理危机的本地化回应。从制度效果上看，韩国的执法实践在一定程度上弥合了两种理念，包括数据处理合法性的判断逻辑，以及对违法训练成果持续效力的终止需求。然而，这种扩张性执法也留下了争议与制度性空白：韩国的做法在操作层面完成了模型层面的风险处置，但在规范结构上仍未明确区分违法行为停止与违法成果法律否定之间的界限。这一实践也未在理论上清晰回答一个更为根本的问题：当模型能力源自违法训练时，其法律效力是否当然随违法行为的确认而

〔59〕 European Commission, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 (Digital Omnibus), COM/2025/837 final.

〔60〕 See Jasmine Park, *South Korea: The First Case Where the Personal Information Protection Act was Applied to an AI System*, FPF (21 May 2021), https://fpf.org/blog/south-korea-the-first-case-where-the-personal-information-protection-act-was-applied-to-an-ai-system/, visited on 6 January 2026.

〔61〕 Ibid.

〔62〕 서인숙, 개인정보위, 개인정보 무단 국외 이전한 카카오페이 · 애플에 총 83억 7,520만 원 과징금 · 과태료 부과, https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do? bbsId=BS074&mCode=C020010000&nttId=10955, 2025 年 11 月 17 日访问。

〔63〕 See Anu Bradford, *The Brussels Effect: How the European Union Rules the World*, Oxford University Press, 2020, pp. 265 - 288.

丧失，抑或仍需独立的规范依据予以否定。这一理论问题恰恰构成中国制度路径需要回应的核心张力。

四、中国法律制度中的潜在落地路径

与美国与欧盟的制度路径相比，中国在模型删除问题上的规范落点呈现出明显差异。美国的“算法没收”主要嵌入消费者保护法框架之中，欧盟呈现 GDPR 与人工智能法并行推进的结构，韩国则呈现出一种过渡性特征，既不同于美国以消费者保护为中心的“算法没收”逻辑，也未依赖独立的人工智能专项立法，而是在个人信息保护法框架内，将违法训练、风险防控与模型结构处置加以联结。中国的制度结构在立法路径上更接近欧盟与韩国模式，即通过统一的个人信息保护立法确立数据处理合法性与行政监管框架，而非依赖消费者保护法进行间接规制。模型删除问题自然首先落入个人信息处理合法性的范畴，而非消费者权益保护或 AI 法的规制领域。

中国当前仍主要停留在以个人信息与数据处理活动为中心的基础治理阶段，对模型作为独立风险载体的制度回应尚未充分展开。在模型层面，现实争议更多集中于生成式模型在版权领域引发的训练与输出问题，〔64〕而就模型在个人信息与数据治理层面所蕴含的同样紧迫甚至更具结构性的风险，尚未形成与之相匹配的分析框架与治理工具。中国当前的人工智能执法实践，仍以模型备案等前置性监管工具为主，并在 2025 年 9 月进一步引入生成式人工智能内容标识义务、〔65〕拟人化 AI 特殊义务〔66〕等。这些制度安排固然有助于提升透明度与可追溯性，但其主要功能仍然集中于风险预防与过程管理，对模型结构性风险缺乏直接回应能力。

中国当前面临的核心困境在于，无论是《个人信息保护法》《数据安全法》《网络安全法》，还是自 2021 年以来网信办等部委发布的若干 AI 相关管理办法，规制对象均以“处理活动”为中心，而非模型这一技术产物本身。删除权、停止处理、下架应用等措施，本质上均属于行为控制型工具，其假定前提是违法风险会随着处理行为的终止而自然消失。单纯从《个人信息保护法》中的删除权推导模型删除规范张力较大，直接移植美国以剥夺违法收益为导向的算法没收逻辑，也难以与中国行政法与财产权保障框架相契合。在这一背景下，模型删除在中国语境中的制度定位尤需谨慎。〔67〕模型删除并非中国法对域外制度的突兀移植，而是现有规则体系在面对深度学习模型结构性风险时可能进一步演化的方向。笔者认为，模型删除不宜被理解为一项“数据主体权利的自然延伸”，亦不宜被塑造为类似美国“算法没收”的惩罚性工具，其更应被理解为一种可以同时纠正持续性和结构性双重风险的行政监管手段。其功能不在于对既往违法行为进行惩

〔64〕 参见蒋舸：《论人工智能生成内容的可版权性：以用户的独创性表达为视角》，载《知识产权》2024 年第 1 期，第 36 页；朱阁等：《人工智能生成的内容（AIGC）受著作权法保护吗》，载《中国法律评论》2024 年第 3 期，第 1 页。

〔65〕 参见《人工智能生成合成内容标识办法》。

〔66〕 参见《国家互联网信息办公室关于〈人工智能拟人化互动服务管理暂行办法（征求意见稿）〉公开征求意见的通知》，国家互联网信息办公室 2025 年 12 月 27 日，https://www.cac.gov.cn/2025-12/27/c_1768571207311996.htm，2026 年 1 月 6 日访问。

〔67〕 See Alessandro Achille et al., *AI Model Disgorgement: Methods and Choices*, 121 Proceedings of the National Academy of Sciences e2307304121 (2024).

戒，而在于阻断违法训练及模型能力持续外溢，从而恢复合法、可控的技术运行秩序。

现行法框架中，除《个人信息保护法》外，并不存在其他更为直接且适配的法律依据。《民法典》虽规定隐私权与个人信息权益，但其规范设计以私法救济为核心，强调个体请求权与侵权责任构造，难以回应模型层面持续性结构风险的问题。《消费者权益保护法》亦主要围绕商品与服务交易关系展开，其规制对象与大模型训练中的数据合法性问题并不完全契合。

中国尚未制定专门的人工智能基本法，相关治理仍处于部门规章与专项规定并行的阶段。近年来，国家网信部门已通过部门规章与规范性文件对算法推荐、深度合成、生成式人工智能等领域作出专门规定。这些专项规范与《个人信息保护法》之间的体系关系尚未在立法层面作出明确界定。规范层级结构上，网信部门出台的专项规定属于行政法规或部门规章，其法律位阶低于全国人大制定的《个人信息保护法》。在法律适用逻辑上，专项规范更多承担具体化与补充性功能，其合法性基础仍需要回溯至上位法授权。在涉及违法训练是否成立、违法状态是否持续以及行政机关是否具有采取结构性处置措施的权限时，仍然需要建立在《个人信息保护法》之上。

下文将对中国个人信息保护法中可能承载模型删除功能的若干规范进路进行系统性梳理与比较，重点考察其法律基础、规范强度与制度限度。分析将围绕删除权的延伸解释、行政机关的宽泛纠正性授权、通过补充立法明确模型删除规则、没收违法所得机制的适用可能性，以及暂停或终止服务等既有监管工具展开，在此基础上形成对中国语境下模型删除可行路径的判断。

（一）删除权的延伸（《个人信息保护法》第 47 条）

《个人信息保护法》中最直观、最容易被首先联想到的制度基础，是第 47 条所确立的个人信息删除权。该条明确赋予个人在特定情形下请求个人信息处理者删除其个人信息的权利，包括处理目的已实现、无法实现或者不再必要，处理行为违反法律、行政法规，或者个人撤回同意等情形。从文本结构上看，删除权以“个人信息”为直接客体，其制度设计显然以传统数据处理场景为背景，假定个人信息以可识别、可定位、可分离的形式存在于数据库或信息系统之中。然而，如上所述，在深度学习模型尤其是大规模生成式模型的语境下，个人信息并非以独立条目或字段形式存在，而是通过高维参数、权重分布与特征关联的方式嵌入模型之中，这一技术现实使得删除权在模型层面的适用天然面临断裂。

试图将删除权直接延伸至模型层面常依赖“功能性等同”的推理路径，即认为当个人信息已不可逆地融入模型并持续影响模型输出时，单纯删除源数据不足以实现“删除”的规范目的。因此有必要通过删除模型本身或其相关训练成果实现对个人信息的实质性保护。〔68〕这一推理在价值层面具有一定吸引力，尤其是在模型记忆、隐私泄露和再识别风险日益凸显的背景下，删除权若止步于数据层面，确有被技术架空的危险。问题是，《个人信息保护法》的删除权并非结果导向的抽象权利，而是高度情境化、以可操作性为前提设计的权利，其行使对象、履行方式与效果评估均以“个人信息处理活动”为中心展开，非以“技术系统的最终状态”为评价基准。

删除权在中国法中的制度定位，始终是一种以恢复个人对信息流转的控制为目标的权利救济

〔68〕 See Cheng-chi Chang, *When AI Remembers Too Much: Reinventing the Right to Be Forgotten for the Generative Age*, 19 Washington Journal of Law, Technology & Arts 22 (2024).

工具，而非针对技术系统本身的结构性纠正机制。〔69〕删除权被触发时，法律所要求的只是“删除个人信息”，而非消除一切可能由该信息引发的衍生影响。将模型删除直接视为删除权的当然延伸，意味着把一项以个人请求为起点、以特定信息为客体的权利，转化为一种足以导致模型整体失效甚至财产性灭失的强制措施。这不仅在规范强度上发生跃迁，也在制度功能上发生了根本转向。若缺乏明确的比例性标准与程序性约束，这种转向难以与删除权原有的权利属性相协调。

（二）宽泛授权（《个人信息保护法》第 61 条第 4 项）

相较于以个人权利为起点的删除权，第 61 条第 4 项所确立的“调查、处理违法个人信息处理活动”的监管权力，在规范结构上更接近模型删除所需的制度形态。该条并未以具体权利类型或具体处理行为为限，而是以“违法个人信息处理活动”这一开放性概念为核心，通过“调查”“处理”两项高度概括的表述，为监管机关保留了广泛的裁量空间。这种以活动为规制对象、以纠正违法状态为目的的授权方式，使其在理论上具备向模型层面延伸的可能性。

从文义和体系解释来看，第 61 条第 4 项并未限定“处理活动”的具体形式，也未将其局限于单次、可分离的行为过程。相反，在算法和模型主导的信息处理场景中，个人信息处理往往体现为一种持续性的技术运作状态，而非孤立的数据操作行为。这一意义上，模型的持续运行、本身即可能构成违法个人信息处理活动的载体或结果。当模型是在违法收集、违法利用或违法整合个人信息的基础上形成，并通过持续部署对外提供服务时，将其视为“违法个人信息处理活动的延续形态”，在规范逻辑上并非不可接受。

正是基于这一理解，第 61 条第 4 项在理论上可以为模型删除提供一种间接的制度锚点。监管机关若认定某一模型的存在和运行，使得违法个人信息处理状态无法通过单纯的数据删除、行为终止或流程整改而消除，则“处理违法个人信息处理活动”这一授权，原则上并不排斥采取更为结构性的纠正措施。与删除权不同，这种路径并非以个体权利请求为出发点，而是以恢复合法处理秩序、消除持续性违法状态为目标。然而，第 61 条第 4 项本质上是一项宽泛授权，其功能在于赋予监管机关应对复杂违法形态的灵活工具，而非为高度侵入性的措施提供当然正当性。如果不加区分地将“处理违法个人信息处理活动”理解为可以当然涵盖模型销毁或永久性删除，容易导致授权外溢，使纠正性权力演变为事实上的惩罚性制裁，引发比例原则、信赖保护以及财产权保障方面的质疑。其合理功能在于为监管机关在面对模型层面的持续性违法风险时提供介入和升级处置措施的法律起点。模型删除若欲在该条款框架下获得正当性，仍需通过严格的必要性判断、替代措施穷尽以及程序性保障加以约束。

（三）补充立法（《个人信息保护法》第 61 条第 5 项）

第 61 条既有授权结构中，第 5 项所设立的兜底性条款，为模型删除在中国法语境下的制度化引入提供了另一条路径。相较于第 61 条第 4 项以“调查、处理违法个人信息处理活动”为核心的概括性授权，第 5 项通过“法律、行政法规规定的其他职责”这一开放性表述，明确预留了通过后续立法或规范性文件补充监管工具的制度空间。

〔69〕 参见程啸：《论〈个人信息保护法〉中的删除权》，载《社会科学辑刊》2022 年第 1 期，第 103 页。

从体系上看，模型删除并不属于《个人信息保护法》制定时所直接预设的监管措施。无论是权利结构、责任配置，还是执法程序，该法均以传统个人信息处理活动为基本假定，未针对算法模型这一复合型技术产物构建独立规则。若仅依赖第 61 条第 4 项的解释扩张，将模型删除纳入“处理违法个人信息处理活动”的当然手段，容易模糊解释与创设之间的界限，亦可能引发对行政机关越权行使兜底权力的质疑。第 5 项所体现的立法预留机制，恰恰为通过补充立法的方式明确模型删除的适用条件、程序和边界，提供了规范基础。

针对技术复杂、影响深远且可能触及财产权和产业发展的监管措施，我国通常倾向于通过专门立法或授权性规则引入，并非完全依赖执法解释。这一立法路径在数据安全、网络安全以及平台治理领域均有先例。这一制度传统下，将模型删除作为一种新型监管工具，通过补充立法予以明确定位，可能符合中国法律体系对监管确定性与可预期性的要求。

通过补充立法引入模型删除机制，可以在规范层面区分不同类型、风险等级的模型删除措施，避免“一刀切”的制度后果。补充立法不仅可以明确模型删除适用的实体要件，例如违法程度、风险持续性以及替代性技术措施是否已穷尽，还可以同步设定程序性保障，包括听证、申辩、复核与救济路径，防止兜底条款被转化为缺乏边界的自由裁量权来源。相较于单纯依赖宽泛授权进行执法扩张，这种经由立法明确的路径有助于平衡监管有效性与创新保护之间的张力。因此，第 61 条第 5 项的制度意义在于为模型删除这一尚未被正面规定的监管工具，提供了合法进入中国法体系的“接口”。

尽管第 61 条第 5 项为模型删除的明确化提供了制度空间，但补充立法并非唯一且当然的路径。首先，补充立法若明确赋予行政机关销毁或强制处置模型结构的权力，在制度效果上可能显著扩大行政干预空间。模型删除涉及对高价值技术资产的处置，其判断往往依赖复杂的技术评估与风险预测。在技术可验证性、因果归属与风险程度认定尚存不确定性时，过度依赖概括性授权容易扩大行政自由裁量幅度。补充立法若缺乏精细化标准与程序保障，可能在强化监管确定性的同时，引入新的权力边界争议。因此，第 61 条第 5 项更宜被理解为一种制度预留空间。

（四）没收违法所得（《个人信息保护法》第 66 条）

第 66 条确立了对违法个人信息处理行为“没收违法所得”的行政处罚机制，这一规定在模型删除讨论中被直觉性地视为潜在规范依据。〔70〕其逻辑在于，若人工智能模型系基于违法收集或使用的个人信息训练而成，且该模型在市场中持续产生经济利益，则通过没收违法所得的方式，似乎可以进一步推导出对模型本身采取处置措施的正当性。然而，深入考察第 66 条的规范结构与制度功能即可发现，将其延伸为模型删除的法律基础面临显著的法理与技术障碍。

首先，没收违法所得是一项典型的结果导向型经济制裁，其规制对象是违法行为所直接或间接取得的财产性利益，而非违法行为所依托的工具或技术本身。第 66 条关注的核心问题在于“违法获利是否应被剥夺”，而不是“违法状态是否需要被结构性消除”。这一点决定了该条款的规范重心在于经济利益的回收与惩戒，而非风险治理或合规修复。模型若被纳入规制视野，更可能被视为一种可能带来收益的生产要素，而非当然等同于违法所得本身。其次，将模型直接界定

〔70〕 参见王青斌：《行政法中的没收违法所得》，载《法学评论》2019 年第 6 期，第 160 页。

为“违法所得”在法理上存在明显张力。人工智能模型的价值通常来源于多重投入的叠加，包括算力、算法设计、工程优化以及合法与非法数据的混合使用，其经济价值难以与特定违法个人信息之间建立一一对应的因果关系。即便可以确认部分训练数据存在违法情形，也难以据此推定模型整体价值完全源于违法行为。若在缺乏明确因果限度和比例判断的情况下，将模型整体视为违法所得并予以没收，容易突破行政处罚中一贯强调的相当性原则，使没收违法所得异化为事实上的财产剥夺。〔71〕 最后，从制度体系协调的角度看，第 66 条与第 61 条所体现的监管逻辑并不相同。前者属于事后处罚性规范，其功能在于对既有违法行为施加经济制裁；后者则侧重于纠正违法状态、防止风险持续。模型删除这一问题上真正需要回应的是模型持续运行所带来的结构性风险，而非企业是否已因违法行为获得不当经济收益。

正因如此，第 66 条可以与其他纠正性措施并行适用，用以剥夺企业因违法个人信息处理所获得的经济利益，但并不能当然推出对模型本身实施销毁、撤除或强制性删除的结论。若试图以没收违法所得为由直接正当化模型删除，不仅会模糊处罚与治理之间的界限，也可能在实践中引发对行政权力过度扩张的质疑。

（五）暂停或终止提供服务（《个人信息保护法》第 66 条）

第 66 条中“对违法处理个人信息的应用程序，责令暂停或者终止提供服务”的规定，在实践中构成了监管机关介入人工智能系统运行的最直接工具之一。与没收违法所得不同，该措施并不以经济制裁为核心，而是以中止违法处理活动、防止风险继续扩散为目的，制度逻辑明显更接近纠正性监管而非惩罚性处分。尤其是面对以应用程序或在线服务形式向公众提供的大模型产品时，该条款在事实上已经成为监管部门能够快速、有效介入的重要法律依据。

从规范结构上看，该措施的规制对象明确指向“应用程序”，其法律效果是暂停或终止对外提供服务，而非对底层技术系统或模型本身作出直接处置。这一设计反映了立法者对技术中立与比例原则的基本考量，即在违法个人信息处理行为尚可通过停止服务加以遏制时，优先选择对外部使用状态进行控制，而非对内部技术结构施加强制性改变。据此，暂停或终止服务更像一种使用层面的封禁，而不是技术层面的删除。暂停或终止服务并不会当然消除模型中既有的个人信息残留，也不必然阻断模型在其他场景中的再利用可能性。模型可能仍被保留在企业内部，用于后续再训练、内部测试，或在完成整改后重新上线。换言之，该措施的核心功能在于终止违法处理活动的现实表现形式（实践中被描述为“下架”“封禁”“停用”），而非消除模型作为风险载体本身的存在状态，其法律效果仍停留在市场可用性层面，未触及模型的存续与技术结构。

因此，性质上“责令暂停或者终止提供服务”构成了一种重要的“事实前置措施”：当模型因其训练来源、运行方式或输出风险而被认定存在持续性违法状态时，暂停或终止服务往往是监管机关采取的第一步处置方式。如果在服务终止后，通过数据删除、流程整改或技术修复仍无法消除违法处理风险，单纯依赖服务层面的封禁将难以实现个人信息保护法所要求的实质合规。

（六）反思

在中国现行法律文化与制度结构中，将模型删除理解为删除权的自然延伸或惩罚性没收工具

〔71〕 参见杨登峰、李晴：《行政处罚中比例原则与过罚相当原则的关系之辨》，载《交大法学》2017 年第 4 期，第 9 页。

存在规范张力。首先，中国个人信息保护制度虽借鉴欧盟数据保护理念，但整体仍以行政监管为主轴，而非以个体私权扩张为核心。删除权在体系定位上属于人格权益的具体实现方式，其制度目的在于恢复个体对特定信息条目的控制，而非重塑技术结构或消除系统性风险。若将模型删除直接纳入删除权体系，不仅突破删除权的对象边界，也可能导致权利逻辑向技术结构层面过度扩张，从而破坏制度均衡。在私权逻辑层面，将模型删除解释为删除权的技术延伸缺乏充分的体系支撑。其次，模型作为企业投入算力、数据与研发资源形成的技术成果，具有明显的财产利益属性。若将模型删除理解为类似美国“算法没收”的惩罚性剥夺措施，其性质将接近行政没收或行政强制执行。对财产利益的剥夺须有明确法律授权，并符合比例原则与程序保障要求。尽管融入了“没收违法所得”的惩罚机制，《个人信息保护法》并未明确赋予行政机关对模型成果进行没收或销毁的处罚权。简单移植美国式算法没收逻辑可能会与基本行政法原则产生冲突。因此，将模型删除定位为惩罚性剥夺工具并不具备稳固的规范基础。

中国人工智能治理呈现出鲜明的风险预防与秩序维护导向。在此结构下，笔者认为模型删除若被理解为风险消除型纠正措施，更为契合既有制度逻辑。《个人信息保护法》第 61 条第 4 项所规定的监督管理权限具有明显的开放性与弹性空间。该条款虽未具体列举技术层面的处置形式，但其规范目的在于确保个人信息处理活动处于可控、合规与风险可防状态。值得追问的是，模型删除在中国语境下的功能是否仅限于违法状态消除与合规恢复，抑或还包括终止违法训练成果继续发挥效力。从体系解释角度看，当违法训练已转化为模型结构性风险时，第 61 条所承载的风险防控职责并不当然止于行为层面的停止命令，而应涵盖对持续风险载体的结构性处置。

在大模型训练语境下，违法性往往并不以原始数据条目的存续为前提，而是通过训练过程固化为模型参数结构与能力表现。当训练数据的取得或使用缺乏合法基础，违法性便可能从行为层面转化为技术结构层面的持续状态。此时，单纯删除源数据或责令停止特定处理活动，可能仅能暂时中断行为链条，但无法消除由违法训练形成的结构性违法状态。因此，本文进一步主张，第 61 条第 4 项的弹性授权，应通过体系解释具体化为多层次的模型处置路径：在风险尚可修复的情形下，可以通过技术重配（如再训练或参数调整）实现结构性纠正；在风险已具外溢可能的情形下，可以采取停止部署或市场撤回措施以即时阻断风险；在违法训练与模型结构高度不可分离、且替代性修复措施不足以消除风险时，则可以结构性删除作为最后手段。三种处置形式并非惩罚梯度，而是风险治理工具的差异化展开。通过对第 61 条第 4 项进行风险导向的体系解释，可以在不突破现行法边界的前提下，为模型层面的风险治理提供充分的规范空间，同时避免与行政处罚法定原则及财产权保障原则发生直接冲突。

讨论模型删除制度的适用边界时，还有必要重视特定领域算法与通用大模型（或基础模型）之间的结构差异。本文涉及过往模型删除案例多涉及特定用途算法，通常围绕相对封闭的数据集与明确功能目标进行训练，模型能力与特定数据之间的对应关系较为清晰。违法数据对模型结构的影响在技术上更易识别，也更可能集中于特定功能模块。在此情形下，停止部署或整体终止相关算法，往往能够较为直接地实现风险阻断与违法状态的终止。相比之下，通用大模型或基础模型以海量、多源数据为训练基础，其能力形成呈现出高度分布式与统计性特征。违法数据的影响可能弥散于参数空间的多个层级，而非集中于单一模块或功能输出。违法训练对模型能力的塑造

因此更为复杂，亦更难通过简单的整体终止实现精确治理。若违法因素能够在技术上被识别与隔离，通过再训练、去学习或参数重构等方式进行结构性修复，更符合比例原则与技术合理性要求。唯有在违法训练与模型整体能力不可分离、且技术修复措施不足以消除持续风险的情形下，结构性删除方可作为最后手段加以适用。

五、结　语

在机器学习语境下，训练数据的违法性已难以通过传统的数据治理工具在规模上加以消解。随着模型规模扩大、训练数据来源高度异质化，违法数据往往不再以孤立、可识别的形式存在，而是被压缩、嵌入并固化为模型能力的一部分。单纯依赖数据删除、停止处理或个案化权利救济，已不足以实现法律所要求的风险消除目标。若法律只能要求停止违法处理，却无法触及因违法训练而持续发挥作用的模型能力，则违法状态在规范意义上便难以真正终结。模型删除因此并非对既有数据保护逻辑的激进突破，而是其在模型时代得以自洽运作所必需的制度补充。

在中国法语境中，这一问题尤为突出。一方面，中国人工智能监管长期依赖行为控制型工具，其优势在于灵活与可执行性，但其局限在于难以应对模型层面的结构性残留风险。另一方面，在生成式人工智能快速发展的现实背景下，大规模训练数据的来源与合规性难以逐一核验。若缺乏模型层面的处置工具，监管机关将面临两难：要么过度依赖事前审查，要么在风险已结构化为模型能力后缺乏有效回应手段。

比较美国、欧盟与韩国制度路径可以看到，不同法域为模型层面处置提供了不同的制度载体。美国将算法处置嵌入消费者保护逻辑，欧盟在数据保护与人工智能专项立法中并行推进，韩国则在个人信息保护法执法实践中实现操作层面的弥合。相较而言，中国在尚未制定专门人工智能基本法的前提下，最为直接且具体系基础的法律锚点仍然是《个人信息保护法》。通过对第 61 条第 4 项监督与风险防控职责的体系解释，可以为模型层面的风险治理提供规范空间，而无须将模型删除构造为私权扩张或惩罚性没收工具。

在此框架下，模型删除应被理解为一种以终止违法状态为目标的纠正性监管手段。具体而言，模型删除可设置三种不同强度的处置形态：其一，市场撤回式模型删除，即通过停止部署、下架应用或限制特定场景使用，使模型在法律与现实层面“不可运行”；其二，技术重配式模型删除，即通过再训练、参数调整或机器去学习等方式削弱或消除违法数据对模型能力的影响，在确保风险可控的前提下实现结构性修复；其三，物理损毁式模型删除，即在违法训练与模型整体能力高度不可分离、且技术修复措施不足以消除持续风险的情形下，对模型进行终局性销毁。

三种处置形态的适用应以违法性质、风险外溢程度与技术可验证性为核心判断因素，并在比例原则与程序保障的约束下加以实施。〔72〕尤其是在区分特定领域算法与通用大模型时，应避免机械适用统一标准：对于结构封闭、功能明确的特定用途算法，市场撤回式或物理损毁式措施可

〔72〕 See Tobias Mahler, *Between Risk Management and Proportionality: The Risk-based Approach in the EU's Artificial Intelligence Act Proposal*, Nordic Yearbook of Law and Informatics 247 (2022).

能更具可行性；而对于能力高度泛化、参数结构复杂的基础模型，则更有必要优先评估技术重配式路径，以在确保风险消除的同时维护技术与产业稳定。

在中国制度语境中，模型删除的价值并不在于威慑或制裁，而在于提供一个能够真正触及模型结构、终止违法状态持续性的法律接口。通过在《个人信息保护法》第61条第4项的风险防控框架下明确三种处置形态及其适用边界，可以在不突破现行法体系的前提下，弥合行为违法与结构违法之间的断裂，并为模型时代的数据治理建立闭合的规范逻辑链条。

Abstract: As artificial intelligence models scale in size and complexity, the unlawfulness of training data can no longer be effectively addressed through traditional data deletion or behaviour-based enforcement tools. In deep learning systems, unlawful data use may become embedded in model parameters and capabilities, such that risks persist even after the original data is deleted or processing activities cease. In the United States, the European Union and Korea, model deletion has been framed across jurisdictions—as algorithmic disgorgement under consumer protection law, as an extension of data protection rights, or as a corrective regulatory measure aimed at terminating structural risk. Within China's existing legal framework, prior to the enactment of a comprehensive AI statute, the interpretive space within Article 61 (4) of the PIPL provides a viable foundation for addressing model-level structural illegality. Rather than constructing model deletion as an extension of private rights or as a punitive confiscation tool, it is advised to conceptualise it as a corrective mechanism aimed at terminating continuing unlawful states. By reframing model deletion as a mechanism for closing the structural gap between unlawful conduct and persistent model capacity, it is possible to offer a doctrinally grounded pathway for China's evolving regulatory architecture.

Key Words: model deletion, ai regulation, algorithmic disgorgement, model withdrawal/recall, machine unlearning/retraining

(责任编辑：张金平)