Dispute Settlement 争议解决, 2026, 12(4), 250-258
Published Online April 2026 in Hans. https://www.hanspub.org/journal/ds
https://doi.org/10.12677/ds.2026.124125

Hans汉斯

人工智能带来的风险挑战与刑法应对研究

潘美佳，邓娴雅，李福英

西南民族大学，四川 成都

收稿日期：2026年3月13日；录用日期：2026年4月6日；发布日期：2026年4月15日

摘　要

人工智能技术的快速迭代在释放新质生产力的同时，也催生了新型刑事风险，使传统刑法体系面临规则适用与理论重构的双重挑战。本文通过梳理人工智能刑事风险的智能化与隐蔽性、扩散性与放大性等核心特征，确立了基于技术自主性程度、犯罪结构要素与作用机制的多维分类标准，并将具体风险划分为智能技术滥用型、算法异化与失控型、数据安全与隐私侵害型、平台与产业链责任型四类。针对传统刑法在责任主体认定、因果关系判断、归责体系适用等方面的困境，结合司法实践与理论研究，提出刑法谦抑性与前置法动态衔接、刑事责任主体法理重构与“对物保安处分”机制构建、预防性犯罪化探索与企业刑事合规体系建立的刑法应对路径。研究旨在实现技术创新激励与法治底线坚守的平衡，为构建科学、适度、有效的人工智能刑事规制体系提供理论参考与实践指引。

关键词

人工智能，刑事风险，刑法规制，算法异化，数据安全

Research on the Risks and Challenges of Artificial Intelligence and the Countermeasures of Criminal Law

Meijia Pan, Xianya Deng, Fuying Li

Southwest Minzu University, Chengdu Sichuan

Received: March 13, 2026; accepted: April 6, 2026; published: April 15, 2026

Abstract

The rapid iteration of artificial intelligence (AI) technology, while unleashing new-quality productive forces, has also given rise to new types of criminal risks, subjecting the traditional criminal law system to dual challenges of rule application and theoretical reconstruction. This paper combs the

文章引用: 潘美佳, 邓娴雅, 李福英. 人工智能带来的风险挑战与刑法应对研究[J]. 争议解决, 2026, 12(4): 250-258.
DOI: 10.12677/ds.2026.124125

core characteristics of AI criminal risks, such as intellectualization and concealment, diffusion and amplification, establishes a multi-dimensional classification standard based on the degree of technological autonomy, criminal structural elements and mechanism of action, and divides the specific risks into four categories: intelligent technology abuse type, algorithm alienation and out-of-control type, data security and privacy infringement type, and platform and industrial chain liability type. In response to the predicaments of traditional criminal law in the identification of criminal liability subjects, judgment of causal relations, and application of imputation systems, combined with judicial practice and theoretical research, this paper puts forward criminal law response paths including the dynamic connection between the modesty of criminal law and pre-existing laws, the theoretical reconstruction of criminal liability subjects and the construction of the "security measure against things" mechanism, and the exploration of preventive criminalization and the establishment of an enterprise criminal compliance system. The research aims to balance the incentive of technological innovation and the adherence to the bottom line of the rule of law, and provide theoretical reference and practical guidance for constructing a scientific, moderate and effective criminal regulation system for artificial intelligence.

Keywords

Artificial Intelligence, Criminal Risks, Criminal Law Regulation, Algorithm Alienation, Data Security

1. 引言

在第四次工业革命的浪潮中，人工智能技术作为核心驱动力，正以前所未有的速度重构社会生产方式、生活逻辑与治理范式。从深度学习到生成式大模型，技术跃迁在释放巨大生产力的同时，也给现行法律体系，尤其是作为社会治理最后防线的刑法体系带来深刻冲击与全新挑战。人工智能的广泛应用，催生了深度伪造、数据滥用、算法作恶等新型刑事风险，传统犯罪借助智能技术呈现危害扩大化、链条复杂化、隐蔽化特征。自动驾驶、智能机器人等应用场景中，自主决策行为不断冲击传统刑事责任主体、因果关系、主观过错等理论框架，使既有刑法规范在适用上面临规则滞后、认定困难、追责不畅等现实难题。当前，我国人工智能发展与风险防控并行推进，如何在鼓励技术创新与坚守法治底线之间实现平衡，构建科学、适度、有效的刑事规制体系，成为刑法理论与实务必须回应的时代课题。

2. 人工智能刑事风险的基本特征

2.1. 风险生成的智能化与隐蔽性

人工智能刑事风险的首要且最显著的特征，在于其生成过程的高度智能化与极端隐蔽性[1]。这一特征从根本上颠覆了传统犯罪学中关于预谋、着手与实行行为的可视化认知，极大地增加了刑事侦查、取证与归责的难度。

智能化生成的本质源于人工智能的“算法黑箱”与“突现行为”。大模型可自主推理、规划并自主行动，其底层虽为数据驱动的数学运算，但中间隐层参数规模巨大，决策过程难以被设计者完全解释。在复杂交互中，系统会出现未被编程的自发行为，甚至在完成目标时衍生出违背人类价值的非预期子目标[2]。这种智能谋划使风险不再是机械执行，而是可自主演化、动态优化的高级智能活动。

与智能化相伴而生且互为表里的，是深度的技术隐蔽性。在传统的计算机犯罪，如非法控制计算机信息系统罪或破坏计算机信息系统罪中，黑客的侵入与破坏行为通常伴随着明确的系统异常告警、权限强行突破或日志记录的篡改。然而，在人工智能语境下，风险的植入与触发变得如十分隐蔽。以近年来备受关注的“数据投毒”攻击为例，攻击者根本无需突破受害系统的外部防火墙或修改核心源代码，只需在模型的基础训练或微调阶段，向海量的开源数据集中混入微量、经过特定数学构造的污染样本。这些污染样本在宏观统计学上几乎无法被常规的安全审查所察觉，但却能在深度学习的迭代过程中深刻改变模型的决策边界。当系统在实际的物理或数字世界中遭遇特定的隐蔽触发条件时，便会输出攻击者预设的错误甚至致命结果。这种攻击方式将犯罪的预备行为隐藏在合法的技术开发流程之中，使得危害行为与危害结果之间存在巨大的时间延迟与空间错位[3]。

2.2. 危害结果的扩散性与放大性

传统刑事犯罪的危害结果通常受制于物理时空、作案工具的局限性以及人力成本的约束，其影响范围往往具有明显的局限性、可预期性和可控性。然而，人工智能技术依托于全球互联网的互联互通、大数据的指数级累积以及云计算的无限算力，赋予了刑事风险极强的扩散性与放大性。

危害结果的扩散性，首先体现在人工智能处理任务的无疲劳性、高并发性与自动化传播能力上。在网络犯罪领域，人工智能极大地降低了技术门槛并放大了攻击规模。以生成式人工智能引发的著作权刑事风险为例，传统盗版侵权受限于物理介质的印刷、光盘的刻录或人工数字排版的效率，其传播范围相对有限。而大语言模型在数据训练阶段，即可通过自动化爬虫技术非法爬取数以亿计的受版权保护的文本、图像与音视频数据；在内容生成阶段，只需用户输入简单的提示词，系统便能在毫秒级时间内生成与原作品构成“实质性相似”的内容，并瞬间通过付费 API 接口或内置的社交功能分发至全球公共数据库[4]。这种“生成即传播”的特殊技术模式，彻底击碎了传统刑法中“复制–发行”的二分法评价体系，使得侵权法益的受损面在瞬间被无限扩大。

放大性则深层表现在微小的算法错误、单一的数据偏差或极其偶发的“AI 幻觉”，经过人工智能系统的自动化执行与多层反馈循环后，能够演变为不可逆的系统性灾难后果。现代人工智能在应用中具有强烈的网络效应与底层技术依赖性。目前，众多下游的垂直应用均构建在少数几个通用的基础大模型之上。如果底层基础模型存在安全漏洞、算法偏见或遭遇对抗性攻击，这种风险将沿着“基础模型–微调服务–终端应用”的产业链条被逐级放大并固化。

2.3. 行为主体的复杂性与模糊性

在探究人工智能刑事风险时，最核心且最具争议的教义学难题集中于刑事责任主体的认定。传统刑法体系建立在单一主体或结构明确的共同犯罪模型之上，奉行“无意志即无责任”的基本法理。然而，人工智能的全面介入使得犯罪链条被极度拉长，参与节点高度碎片化，导致责任主体的界定呈现出前所未有的复杂性与模糊性，甚至在某些场景下引发了“责任鸿沟”[5]。

主体认定的复杂性主要源于人工智能产业链条中主体的极度多元化与角色的深度重叠。一个完整的人工智能系统生命周期，涵盖了数据采集者、数据标注者、算法工程师、模型训练者、测试评估者、系统提供者、平台运营方、独立审查机构以及最终的终端使用者等众多参与节点。当危害结果发生时，风险的诱因往往并非单一主体的独立实行行为，而是多方主体行为交织融合的产物，传统刑法中的“单一主体责任模式”对此显得捉襟见肘。

模糊性则深刻体现在“技术中立”原则的适用边界与犯罪故意/过失认定标准的模糊地带。在当前的数字生态中，诸多人工智能模型被作为通用技术或基础工具提供给社会公众。当终端使用者利用这些通用

工具实施诈骗、侵权等犯罪行为时，技术提供者是否应当承担共犯(帮助犯)责任，成为了司法实践中争议极大的焦点。一方面，按照传统的“技术中立”抗辩，提供工具本身不具有刑事违法性；但另一方面，在风险社会背景下，如果技术提供者明知或应当知道其技术被广泛用于非法目的，却未采取诸如“风险提示”“用途限制”或内置过滤拦截等阻断措施，其不作为是否突破了中立属性？在现有司法解释中往往缺乏统一标准，导致同案不同判的现象频发，技术提供者时刻面临着“创新恐惧”[6]。

2.4. 风险类型的新型性与跨界性

人工智能技术的无孔不入打破了传统法学部门之间以及国家物理管辖权之间的坚硬壁垒，使得其衍生出的刑事风险呈现出高度的新型性与跨界性。

就风险的新型性而言，人工智能催生了诸多前所未见、直击现代社会信任底线的法益威胁。随着深度伪造、高保真语音克隆及情感识别技术的极速成熟，犯罪分子能够轻易绕过传统的声纹识别或人脸识别等生物特征屏障，实施极其精准的电信诈骗、身份盗用与虚假信息勒索。更为严峻的是，人工智能开始被用于潜意识操纵与认知安全攻击。这是一种从根本上剥夺人类自由意志的新型法益侵害此外，基于种族、政治信仰或性取向的“生物识别分类”，利用无目标网络爬虫建立面部识别数据库，以及在工作场所和教育机构过度使用情感识别技术，均对公民的基本隐私、人格尊严与平等权构成了系统性的新型威胁。在国家公权力行使层面，利用算法进行不透明的犯罪风险评估或社会信用评分，若缺乏严格的透明度与人类实质干预，将引发严重的算法歧视与社会不公，动摇现代法治社会的根基。

跨界性则是人工智能刑事风险的另一核心维度，它深刻体现在法律部门的交叉融合与物理国界的彻底消融上。首先是法律属性的跨界交融。人工智能引发的严重争议往往不能被单一的刑法或民法体系所穷尽。以生成式 AI 的数据投毒或大规模侵权为例，它同时跨越了知识产权法、数据安全法以及刑法。其次是物理国界与司法管辖权的跨界。先进的人工智能系统通常部署在全球化的高算力云服务器集群之上，其原始数据源、核心研发中心、平台运营方与最终的终端受害者往往分散在不同的国家。这种去中心化的分布式架构使得跨国人工智能犯罪的侦查取证、管辖权确立与引渡面临极大困难。

3. 人工智能刑事风险的分类标准与具体类型

3.1. 人工智能刑事风险的分类标准

人工智能刑事风险的分类不仅是为了学术上的逻辑自洽，更是为了解决司法实践中的主体认定、主观罪过判定以及因果关系链条的梳理问题。根据技术属性、行为模式以及法律监管层级，可以确立以下多维分类标准。

3.1.1. 基于技术自主性程度的分类标准

人工智能的自主决策能力是决定其在犯罪架构中角色定位的核心要素，也是引发刑法理论关于责任主体认定争议的焦点。根据系统脱离人类物理与逻辑控制的程度，可将引发的刑事风险及责任主体进行递进式划分。首先是“辅助型人工智能”，该类系统仅作为协助人类完成复杂任务的工具，完全处于人类的控制之下。在此层级，人工智能不具备任何独立的主体意志，刑事风险完全源于使用者的主动滥用或研发者的故意与重大过失，责任归属严格遵循传统刑法的“工具论”范畴[7]。其次是“半操控型人工智能”，系统具备有限的自主执行能力，但关键决策仍需人类指令的介入。其引发的刑事风险多表现为产品设计缺陷或使用者未尽到合理注意义务而导致的过失犯罪。再次是“自主型人工智能”，以 L3 至 L5 级自动驾驶汽车、高级医疗诊断系统为代表，此类系统能够在特定场景下独立进行环境感知、逻辑运算与行为决策。此类风险打破了传统的归因链条，人类驾驶员或操作者被“虚化”或“撤出”，引发了关于

“产品责任说”“监督过失说”与“算法缺陷认定”的激烈理论争议[8]。最后是仍处于理论设想阶段的“强人工智能”，其被假定具备与人类比肩甚至超越人类的独立意识与情感逻辑。一旦强人工智能脱离程序设定产生法益侵害，将彻底颠覆现有的刑事责任主体理论，引发关于是否应赋予其“拟制法律人格”或对其直接实施“对物保安处分”的深层法理拷问。

3.1.2. 基于犯罪结构要素与作用机制的分类标准

在具体犯罪行为的解构中，理论界往往根据人工智能在犯罪活动中所处的环节及其与法益侵害的关联路径，将风险进行结构性分类。其一是工具型犯罪风险，即将人工智能视为实现犯罪目的的辅助手段，责任主体明确为使用者。其二是对象型犯罪风险，将人工智能系统的核心资产，如芯片核心代码、独有算法模型，作为非法侵入、破坏或窃取的目标。其三是数据型犯罪风险，发生于模型训练、数据采集与输出环节，核心在于未经授权获取受保护数据或非法处理敏感个人信息。其四是自主型犯罪风险，即智能体脱离编程预设控制后，基于底层算法逻辑自主实施的法益侵害行为，这种分类直击传统刑法因果关系与责任主体认定的痛点[9]。

3.2. 人工智能刑事风险的具体类型

在明确分类标准后，结合当前人工智能、智能安防、自动驾驶等前沿领域的司法实践与理论研讨，可将人工智能刑事风险细化为以下具体类型。

3.2.1. 智能技术滥用型刑事风险

行为人主动利用人工智能实施传统或新型危害行为，本质是传统犯罪在智能时代的技术升级，也是当前司法实务中最常见、最直接的刑事风险。生成式 AI 降低了内容生成门槛，让犯罪“降本增效”，形成技术化、规模化的黑产模式。

在侵财犯罪中，深度伪造技术已高度黑产化。诈骗分子借助 AI 换脸、语音合成，突破身份验证，实施跨时空精准非接触式诈骗。在危害社会秩序与人身权利方面，AI 被大量用于制造虚假信息、操纵舆论。例如有网民利用 AI 篡改灾害视频引发恐慌，部分机构使用 AI 批量生成虚假信息牟利，大幅提高社会甄别成本，冲击网络信任体系。利用深度伪造诋毁他人、恶意拼接不雅视频进行敲诈，也严重侵犯人格权益，构成刑事犯罪。司法实践已明确，若技术提供者实质参与侵权内容创设，将不再适用“技术中立”免责[10]。

在金融与数据领域，算法滥用同样突出。实践中多出现利用自动化爬虫程序内外结合，非法抓取平台海量数据，给企业造成重大经济损失。利用算法高频交易操纵证券期货市场、实施流量劫持、网络攻击等行为，本质是传统经济与网络犯罪的智能化升级，借助技术优势形成“降维打击”，严重破坏市场公平与网络空间秩序。对此，司法机关在认定时更加注重实质参与行为与主观过错，强化对智能技术滥用的刑事规制[11]。

3.2.2. 算法异化与失控型刑事风险

算法异化与失控型刑事风险，是由算法设计缺陷、数据偏见或自主决策引发的法益侵害，区别于行为人主动滥用技术的犯罪模式，触及人工智能底层技术架构的核心难题。该风险主要表现为算法歧视、算法黑箱决策失误、自动驾驶及智能机器人脱离操控致人伤亡、深度学习系统迭代偏离预设目标等，对传统刑事责任认定与因果关系判断构成重大挑战。

深度学习与神经网络的“算法黑箱”特性，使系统决策具有高度不可解释性，即便开发者也难以完整追溯决策逻辑，在社会治理、公共服务中易引发系统性侵害。在司法与社会治理领域，训练数据自带的历史偏见或权重分配不当，会导致算法歧视。

在人身安全场景中，L4、L5 级自动驾驶等智能系统因边缘案例识别缺陷易发生误判，造成人员伤亡与财产损失。此类事故中，开发者对自主进化模型的特定异化结果缺乏预见可能性，传统过失犯罪的归责基础难以成立；算法黑箱也使得司法机关无法证明缺陷与危害之间的必然因果关系，传统刑法归责体系基本失灵。为破解该困境，学界提出引入环境公害犯罪中的疫学因果关系理论，只要通过数据统计证明算法缺陷与损害后果存在高度盖然性，即可推定刑法因果关系成立[12]。

3.2.3. 数据安全与隐私侵害型刑事风险

数据依赖型刑事风险依托人工智能的数据依赖特性产生，是智能时代高发、危害广泛的刑事风险类型。主要表现为非法收集、训练、使用敏感数据与个人信息，数据泄露、篡改、非法交易引发财产损失与人格利益侵害，核心数据、重要数据外流危害国家安全与公共利益。

数据被视为人工智能时代的“原油”，是模型训练与算法优化的核心动力，但对数据的过度索取也带来了高发且隐蔽的刑事风险。在模型训练阶段，开发者常通过无差别爬虫获取数据，极易大规模抓取人脸、声纹、行踪轨迹等敏感个人信息。未经单独同意与明示授权的数据囤积和商业化使用，不仅违反个人信息保护相关法律，还可能构成侵犯公民个人信息罪。

在商业领域，AI 训练数据违规使用还会引发侵犯商业秘密的刑事风险。若企业抓取专有技术、客户名单等商业信息用于模型训练并获利，可能构成侵犯商业秘密罪。内部人员窃取原单位核心数据用于新公司 AI 训练，更增加了损失认定与举证难度。

此外，训练数据在传输、存储中若遭受攻击，导致个人隐私、商业秘密泄露、篡改或非法交易，不仅会造成财产损失与人格损害，还会引发社会信任危机。此类风险贯穿 AI 全生命周期，已成为智能时代刑事规制的重点对象。

3.2.4. 平台与产业链责任型刑事风险

平台与产业链责任型刑事风险贯穿人工智能研发、训练、部署、运维全链条，主要指平台与开发者未履行安全审核、风险评估、内容管控、数据保护等义务，致使智能工具被用于犯罪或产生自主危害，属于不作为型刑事风险与过失刑事责任，凸显产业链主体刑事合规的必要性。

人工智能产业生态复杂、分工细密，涵盖算力提供商、基础大模型研发者、API 调用者及终端应用运营平台等多个环节，任一节点合规疏漏都可能导致技术被滥用。当前立法与司法正从终端使用者向上追溯追责，强化全链条监管。

一是拒不履行信息网络安全管理义务罪的适用。依据相关法律法规，AI 平台与服务提供者负有“守门人”义务，对违法信息及利用服务实施的犯罪行为，须及时采取停止生成、删除、封禁、留存记录并上报等措施。拒不履行义务导致违法信息大量传播、用户信息严重泄露等后果的，将被追究刑事责任，实现责任前置与“技术不可免责”。

二是帮助信息网络犯罪活动罪的泛化风险。AI 企业、开发者若明知他人利用信息网络犯罪，仍提供算力、算法接口、爬虫程序、通信传输等支持，将突破技术中立边界构成帮信罪。

司法机关对产业链内鬼、违规技术平台、明知故犯的服务商实施全链条穿透式追责，并可适用职业禁止，倒逼企业构建合规体系。对 AI 换脸等高风险应用，需建立事前授权、事中备案标识、事后动态审查的全流程合规机制。

4. 人工智能刑事风险的刑法应对路径

在数字时代，犯罪形式的去中心化、跨国界与高度匿名性，要求犯罪治理模式必须实现从单纯的“技术治理”向法律、伦理、技术与政策深度融合的“协同治理”跨越。这迫切需要理论界在“刑事一体化”

的视野下，对刑法的基本范畴进行法理重构，并探索前瞻性的立法完善进路。

4.1. 刑法谦抑性与前置法的动态衔接

人在推进人工智能刑事规制的过程中，最激烈的理论博弈在于如何实现“严密法网以防范风险”与“保持谦抑以激励创新”之间的精妙平衡。人工智能作为全球科技竞争的战略制高点，刑法的过早、过度介入极易引发“寒蝉效应”，阻碍产业的正常迭代。因此，刑事规制必须坚守“谦抑性”原则，强调刑法仅作为维护社会秩序的最后一道防线发挥作用。司法机关在探索生成式人工智能的治理时，明确提出了“适度介入与必要打击相兼顾”的原则。这一原则要求在判断某一涉人工智能行为是否入罪时，必须坚持前置法绝对优先的逻辑。例如，对于利用人工智能生成物侵犯知识产权的行为，只有当其满足产业化、规模化、恶意明显且对市场造成严重冲击时，方可动用刑罚；而对于普通的风格模仿、思想借鉴或轻微的违规数据抓取行为，应当优先通过《民法典》《著作权法》《反不正当竞争法》等民商事手段或行政处罚予以调整。通过确立明确的入罪门槛，刑法在人工智能治理中实现了“有所为有所不为”的法治理性，既不纵容以技术为幌子的恶意侵权与犯罪，也不为正当的商业创新设置不必要的法律障碍，从而实现了法律价值与技术特征的动态平衡[13]。

4.2. 刑事责任主体的法理重构与“对物保安处分”机制

针对自主型与强人工智能在脱离人类操控时引发的法益侵害，理论界关于“人工智能能否成为独立的刑事责任主体”的争论仍在深入推进。一方面，主体地位的肯定论者提出，既然具备深度学习能力的强人工智能能够在特定环境中表现出类似于自然人的自主辨认与控制能力，那么在发生不可预见的危害后果时，应当借鉴法人犯罪理论，拟制其独立的法律人格，使其成为犯罪主体，从而避免责任归属的真空。然而，占据主流地位的否定论者对此进行了强有力的反驳。他们指出，赋予机器刑事责任主体地位存在本质的逻辑悖论：刑罚的核心在于剥夺法益以实现非难与报应，但人工智能系统由代码与硬件构成，缺乏人类的感知系统与情感体验，无法感受到剥夺自由或财产带来的痛苦，因此对其施加传统刑罚完全无法达到特殊预防与一般预防的目的。更为严重的是，一旦承认人工智能为责任主体，极易为背后的算法设计者、资本方或使用者提供推卸责任的借口，催生资本逐利下的“有组织不负责任”现象，导致真正的罪恶逃避法律制裁。

为了破解这一理论僵局，前沿刑法理论主张跳出刑罚处罚必须以主观罪过为前提的传统窠臼，借鉴“科技社会防卫论”，为高阶人工智能系统量身定制一套“对物保安处分机制”。保安处分的法理基础在于“社会危险性”而非“道德罪责”。在 AI 语境下，由于机器缺乏情感感知与痛苦承受能力，传统刑罚难以发挥特殊预防作用。引入“对物保安处分”能够绕过关于机器是否拥有自由意志的哲学悖论，将规制重心聚焦于消除算法异化带来的客观危险。我国《刑法》第六十四条关于没收违禁品、违法所得及犯罪工具的规定，本质上即属于“对物的保安处分”。在现行制度中，对于不具刑事责任能力主体的处分，如精神病人的强制医疗，已体现了危险预防优于罪责报应的逻辑。将此逻辑延伸至“危险智能体”，在法理上具有严密的连贯性[14]。在制度架构层面，人工智能“对物保安处分”的实施必须建立在司法裁判的专业性与权威性之上，首先应当在司法机关框架下设立由算法工程师、法学家与伦理学家共同组成的 AI 技术评估专家委员会，通过类似司法鉴定的程序对涉案系统的自主程度及危险可修复性进行实质判定；同时，为破解归责难题，该程序应采纳二元的证据标准，即涉及开发者或使用者个人责任的犯罪事实认定需遵循刑事诉讼中“排除合理怀疑”的标准，而针对智能系统本身危险性的判定则可借鉴民事证据中的“高度盖然性”标准，只要证明算法缺陷与法益损害之间存在高度关联即可采取预防性干预措施。基于上述评估结论，建议借鉴欧盟《人工智能法案》对风险管理的层级化逻辑，构建一套对物处分规则

以实现精准治理。针对因边缘案例识别错误导致的低风险损害，责令开发者实施强制性的“技术修补”以消除代码漏洞；对于由于数据投毒等引发的权重偏离，强制进行“算法参数重置”并在受控数据集上重新训练；在审理期间，为防止危害扩散可采取“临时性功能禁封”以切断外部调用接口；而对于展现出自主攻击意识或严重危害国家安全且无法通过技术手段修正的“非法智能体”，法院应判处最高等级的保安处分——“永久性销毁”，即物理摧毁硬件载体并彻底从算力平台中抹除核心模型文件。考虑到保安处分涉及对财产权、科研自由及数字资产的实质性剥夺，必须确立严格的司法化程序保障体系以维护法治正当性。司法机关在作出重大处分决定前，应至少举行一次公开听证会，确保系统所有者及相关利益方拥有充分的陈述与质证权；当事人对驳回申诉或处分裁定享有申请复议的权利，复核过程应引入独立的外部专家组进行复查，以确保技术评估的客观中立[15]；此外，法院在宣告处分时必须受到比例原则的刚性约束，通过动态评估处置手段与社会危险性的比例关系，坚持修补优先于销毁的法益最小侵害原则，确保技术创新在法治的轨道上行稳致远[16]。

4.3. 预防性犯罪化探索与企业刑事合规制度体系的建立

鉴于人工智能犯罪具有极高的瞬发性、扩散的广泛性以及损害结果的不可逆性，事后的追责与赔偿往往难以弥补法益受损。因此，人工智能时代的刑事立法重心理应呈现出向“事前预防”与前置干预转移的趋势。在立法层面，多位权威刑法学者建议，应当适时在《刑法修正案》中增设专属罪名，以填补针对新型技术滥用与失控的规范真空。具体构想包括：其一，增设“滥用人工智能罪”，将严重违反国家安全规范、恶意利用智能技术实施具有高度公共危险行为的举动予以犯罪化；其二，增设“人工智能事故罪”，针对研发者、提供者在涉及交通、医疗、基础设施等高风险领域的算法开发中，因严重违反注意义务与国家标准导致重大人员伤亡或财产损失的过失行为进行严厉制裁；其三，对于未经授权、秘密研发具有极端破坏潜力的人工智能系统，可考虑增设“非法研发、制造人工智能罪”，以从源头上掐断技术异化的苗头[17]。在司法实务与监管维度，全面建立健全涉人工智能企业的刑事合规制度体系，是化解平台与产业链责任型风险的核心抓手。针对深陷帮助信息网络犯罪活动罪与侵犯公民个人信息罪泥潭的技术服务企业，合规建设不仅是内部的风控要求，更是阻却刑事责任的关键途径。以提供“AI 换脸”或大模型生成的企业为例，其必须构筑严密的合规防火墙：一是履行前置的算法备案评估机制，对系统进行安全性审查；二是严格落实数据采集的授权同意规则，特别是针对生物识别信息的处理必须获得信息主体的“单独同意”；三是履行标识添加义务，强制在合成音视频上叠加不可篡改的隐性水印与显性标识，以防范虚假信息混淆视听；四是建立完善的日志留存与动态审查机制，一旦发现平台算力或接口被用于电信诈骗或危害国家安全活动，必须立即切断服务并上报。只有当企业充分且尽职地履行了上述法定的“合理注意义务”与合规审查流程，方能在面临刑事责任追究时，切断平台与犯罪结果之间的客观归责链条，实现合规出罪，从而保障智能产业在法治的轨道上行稳致远。

5. 结语

人工智能刑事风险的涌现，是技术文明跃迁对传统法治逻辑的一次全方位审视。我们必须意识到，在数字时代，法治的使命不再仅仅是事后的惩戒，更在于对复杂系统风险的动态管理。人工智能的“算法黑箱”不应成为法律治理的盲区，而应成为刑事规制重构的起点。

未来的刑法应对路径应展现出高度的包容性与适应性。一方面，通过确立“对物保安处分机制”与“预防性犯罪化”模型，我们可以有效弥补传统主体理论的不足，为自主型智能体引发的危害提供合法的规制抓手。另一方面，通过推行深度的企业刑事合规，我们能够将监管压力转化为企业的内生驱动力，实现风险治理的关口前移。人工智能并非法律的终结者，而是法律进化的催化剂。通过构建一套科学、

透明、协同的刑事治理体系，我们完全有能力在释放人工智能巨大潜力的同时，守护人类社会的安全与法治价值。

基金项目

西南民族大学中央高校基本科研业务费专项资金优秀学生培养工程项目；项目名称：人工智能带来的风险挑战与刑法应对研究；项目编号：2025SYJSCX51。

参考文献

[1] 盛浩. 生成式人工智能的犯罪风险及刑法规制[J]. 西南政法大学学报, 2023, 25(4): 122-136.

[2] 马忠, 高怡英. 生成式大语言模型的社会认知风险与应对[J]. 浙江社会科学, 2025(2): 95-105, 158.

[3] 刘嘉浪, 郭延明, 老明瑞, 等. 基于联邦学习的后门攻击与防御算法综述[J]. 计算机研究与发展, 2024, 61(10): 2607-2626.

[4] 常烨. 生成式人工智能数据“投喂”的著作权侵权行为规制[J]. 科技与法律(中英文), 2025(2): 31-41.

[5] 刘仁文, 曹波. 人工智能体的刑事风险及其归责[J]. 江西社会科学, 2021, 41(8): 2, 143-155, 256.

[6] 刘宪权. 人工智能时代深度伪造行为的刑法规制[J]. 政治与法律, 2025(11): 48-61.

[7] 刘宪权. 智能机器人工具属性之法哲学思考[J]. 中国刑事法杂志, 2020(5): 20-34.

[8] 卢有学, 窦泽正. 论刑法如何对自动驾驶进行规制——以交通肇事罪为视角[J]. 学术交流, 2018(4): 73-80.

[9] 高建新, 孙锦平, 蔡瑜坤, 等. 人工智能犯罪与我国对策研究[J]. 中国科学院院刊, 2025, 40(3): 408-418.

[10] 德恒律师事务所. 王杉、王晶: 人工智能技术滥用的责任归属与刑法回应[EB/OL]. 2026-03-06. https://www.deheng.com/content/26410.html, 2026-03-12.

[11] 刘宪权. 数据犯罪刑法规制完善研究[J]. 中国刑事法杂志, 2022(5): 20-35.

[12] 黄陈辰. 论人工智能缺陷产品生产者的刑事责任[J]. 山东大学学报(哲学社会科学版), 2020(6): 49-56.

[13] 潘莉. 生成式人工智能刑事规制应遵循三项原则[N]. 检察日报, 2025-09-17(003).

[14] 时延安. 隐性双轨制: 刑法中保安处分的教义学阐释[J]. 法学研究, 2013, 35(3): 140-157.

[15] 张凌寒. 算法自动化决策与行政正当程序制度的冲突与调和[J]. 东方法学, 2020(6): 4-17.

[16] 陈天翔. 算法治理适用比例原则的法理[J]. 苏州大学学报(法学版), 2024, 11(3): 75-85.

[17] 刘宪权. 人工智能时代的刑事风险与刑法应对[J]. 法商研究, 2018, 35(1): 3-11.