DOI: 10.54254/3050-2160/2025.21188

浅析AI污染的法律规制

金斌海

（北京中凯（上海）律师事务所，上海；jbh120015@163.com）

摘　要：由于生成式人工智能（AIGC）技术的广泛应用，致使 AI 生成的大量涉嫌编造、虚假、有害的低质量信息垃圾充斥并堆积于互联网中，形成 AI 污染现象。其危害包括侵害权利、造成社会失序并引发技术异化。探其成因，从生成端看，AI 生成原理天然具有不可解释性，垃圾信息生成无可避免；从传播端看，信息垃圾传播具有低成本性；从积累端看，识别污染信息存在难度，且平台清理动力不足。从当前法律治理模式与现状看看，对 AI 污染的处置规范散见于多部文件，缺乏针对性与体系性。相关措施存在规定模糊、救济滞后和效力不足的困境。因此为深化 AI 污染治理，平衡生成式人工智能技术高速发展与风险管控，可从生成端强化立法留痕，统一标识标准，从传播端加强监管与违法打击，细化平台责任，从积累端健全清理机制，扩充举报渠道，成立专门机构。

关键词：AI 污染；生成式人工智能（AIGC）；法律规制；信息垃圾

2023 年，以 ChatGPT 为代表的生成式人工智能（AIGC）一经发布，便以其颠覆性技术席卷全球，为各行各业的内容生产模式带来了革命性改进。然而，在 AIGC 技术为公众降低内容创作门槛的同时，由 AI 生成的低质量内容不断涌现。凭借算法推荐与网络平台传播力，部分缺乏内涵、观点偏颇、事实失真的 AI 创作正以高速之势侵蚀互联网内容圈的原有生态，信息垃圾被不断塑造并堆积于互联网中，形成了 AI 污染现象。2024 年 4 月 24 日，清华大学新闻与传播学院新媒体研究中心发布的《揭秘 AI 谣言：传播路径与治理策略全解析》研究报告中指出近一年来，经济与企业类 AI 谣言量增速达 99.91%，致使大量低质信息充斥于互联网中。[1]在国外，Cybernews 于 2023 年 12 月 18 日发布的 Report: rise of AI-generated websites turbo-charges spread of misinformation 一文中指出,由生成式人工智能创作的虚假文章网站数量自 2023 年 5 月以来激增 1000%。[2]在数字经济时代下，人类的各项活动已与互联网高度绑定，低质信息的产生与聚集将严重影响公众生活、产业发展与社会进步。法律及相应规制手段作为调整社会关系的有效工具，应当为 AI 污染治理难题提供有益方案，积极发挥指引与监管作用，以实现人工智能产业的良性发展，做到让科技正向回馈公众。

1 AI 污染的危害与成因

AI 污染是指人类使用生成式人工智能技术的过程中，大量经 AI 生成但未经人工核查、高度改进的低质量创作成果被传播并堆砌于互联网与数据库中的现象。该类低质量创作成果通常包含虚构事实、数据失真、煽动引导等内容，具有人工性弱、内涵浅薄、信息贫乏等特征，属于信息垃圾。AIGC 技术本意是作为人类生活工作中的辅佐官，为人类思想创作与生产力提升提供便利，而无节制的技术滥用反致信息垃圾的大量生成，其广泛传播正在不断扩张并掠夺互联网与数据库的原有空间，持续稀释人类原创，其危害具有“涟漪效应”。

首先，信息垃圾将直接侵害权利。一方面，AI 生成内容可能未经授权使用他人文学、艺术素材，其生成结果涉嫌侵犯他人著作权。另一方面，数据收集使用过程中若涉及个人生物信息和其他基本隐私，则侵犯肖像

[1] 2024 年 4 月 24 日，在中国人民大学公共传播研究所与中国人民大学刑事法律科学研究中心联合举办的“AI 谣言现状与治理”专题研讨会上，清华研究团队发布该报告。

[2] Gintaras Radauskas, “Report: rise of AI-generated websites turbo-charges spread of misinformation” Cybernews. Dec 18, 2023. https://cybernews.com/news/ai-generated-websites-misinformation-propaganda/

权、隐私权等。2024 年杭州互联网法院审理的“AI 换脸侵权第一案”中，被告利用开源模型合成他人肖像牟利，被判赔偿 12 万元并公开道歉，凸显 AI 污染的司法认定与治理难题。[3]其次，AI 污染会造成社会失序。生成的低质信息通过算法推荐一旦传播，将可能严重干扰公众信息获取与判断，降低公众思辨能力，有碍公众获取真实信息，塑造错误认知和价值观。尽管部分 AI 看似“上知天文，下晓地理”，但生成的内容却空洞生硬，在人们缺乏批判性思考时，其构建的“知识体系”会导致人们思辨能力退化，陷入认知幻觉，引发公众认知困惑，扭曲公众对现实与科学共识的理解，对年轻一代而言，若认知被“信息垃圾”塑造，后果严重 [1]。再次，部分虚假新闻、不实信息与网络谣言易广泛传播并引发社会恐慌，严重威胁社会安全和稳定。2024 年 1 月，某 MCN 机构利用 AI 批量生成虚假新闻，其中某新闻编造“某地突然响起巨大爆炸声，导致车毁人亡”谣言，并于文字下方配发虚假爆炸图片，引发社会恐慌后被公安机关查处，足见 AI 污染的现实破坏力。[4]最后，即使部分 AI 生成的垃圾内容不具有危害性，这些意义不明、内容寥寥、涵义浅薄的信息若再次成为训练数据，将形成恶性循环，导致 AI 输出质量下降，造成技术异化。这将增强用户的学习成本，浪费用户大量时间，提升 AI 进化难度，最终减缓科技与产业发展步伐。纽约大学的研究团队发现，只需在数据库中输送 0.001%的虚假信息，就足以导致部分人工智能的输出结果产生重大错误，并进而阻碍使用者获取有效信息。[5]

据上述，垃圾信息的积累致使 AI 污染形成，究其原因，可以分别从垃圾信息的生成端、传播端与积累端加以分析。

从垃圾信息生成端来看，一方面，生成式人工智能的决策过程本身缺乏可解释性，其输出结果存在“不良品率”，这使得生成低质内容无可避免。依据“算法黑箱”原则，人类难以全面理解、解释并判断人工智能生成结果的依据。由于其最终生成内容是算法运行的结果，故而天然在一定程度上缺乏真实性、准确性与可靠性，从而必然会有几率产生“信息垃圾”。截止 2025 年 1 月，经测试，OpenAI 的 GPT-3.5 生成结果的虚假概率为 3.5%，GPT-4 为 1.8%。[6]另一方面，低质量内容回流数据库引起的算法偏差导致人工智能更易生成“信息垃圾”，循环加剧污染现象。人工智能算法是基于大量的数据库与网络信息进行训练，但如果训练素材包含偏差、偏见、失准等低质量信息，模型输出的结果可能会在吸收、分析与学习后输出重复、相似、无用等低质内容，且该结果会因再次被纳入数据库，进而影响机器的再学习与再生成。

从垃圾信息传播端看，AIGC 技术的快速发展和 AI 产品涌现系 AI 污染形成的首要原因，它使得生成信息垃圾的难度与成本降低，具有低成本性。普通用户无需专业知识和复杂技能，仅需简单指令即可产出大量“污染源”，使得垃圾信息的产生变得极为容易。其次，发布由 AI 生成的信息垃圾存在可观收益。网络时代，流量与经济利益密切相关。通过大量投放 AI 生成的垃圾信息以霸占网络平台并形成“AI 内容农场”。在算法推荐下，部分用户关注并点击这些垃圾信息，从而为投放内容及发布账号迅速获取流量，并进一步借助平台流量奖励、广告投放、账号销售等途径获取经济利益。而部分平台亦默许“AI 内容农场”存在，甚至主动推荐放大低质内容传播，以换取更多平台流量。最后，发布垃圾信息的后果与法律责任较轻。生成式人工智能技术更

[3] 原告廖某和吴某在刷社交平台时，偶然发现自己精心拍摄并发布的出镜视频竟被某人工智能“换脸”App 软件公司用于制作换脸模板。该公司并未经原告授权同意，设置该模板付费后供用户使用，以此谋取商业利益。原告认为，被告公司侵害了其肖像权及个人信息权益。2024 年 6 月 20 日，北京互联网法院一审开庭宣判被告向原告赔礼道歉，并分别向赔偿廖某和吴某精神损失费 500 元及财产损失 1500 元。

[4] 2024 年 1 月，某网络平台出现一条关于“西安爆炸”的消息称，当月 10 日晚，西安突然响起巨大爆炸声，文字下方还配发所谓爆炸的图片。消息一经发布，很快在网上传播。经当地警方核实，本地并未发生类似事件，初步判定该新闻发布者某 MCN 机构发布的数条信息涉嫌网络造谣。警方对该 MCN 实际控制人王某某进行约谈，得知该热点新闻文本于图片均由 AI 生成，全程无需人工参与。后警方认为王某某行为构成传播谣言，虚构事实，扰乱公共秩序，依法对其处以行政拘留 5 日，责令该 MCN 机构停业整改。

[5] 2025 年 1 月，纽约大学研究团队在对一个名为“The Pile”的人工智能模型进行实验，他们在该人工智能数据库中故意加入了 150000 篇由 AI 生成的医疗虚假文章，尽管该文件量仅占原始数据库 0.001%的原始内容，但加入后的模型生成结果错误率增加 4.8%。然而这个过程的成本极其低廉，仅花费了 5 美元。

[6] 2025 年 1 月，Vectara 公司发布其测试后的 AI 幻觉率排行榜。AI 幻觉率是指 AI 生成内容存在编造或与事实不符的虚假信息的情形。该测试结果指出所有 AI 模型均存在捏造事实、编造不存在信息的的情况，其中最差模型输出结果的幻觉率高达 30%。截止至 2025 年 1 月 5 日，OpenAI 的 GPT-3.5 的 3.5%，GPT-4 为 1.8%，o1-mini 为 1.4%。

新周期较短，但与之相应的监管技术、法律规范以及行业标准等难以随时跟进，对 AI 污染的追责缺少准确标准，实践规制存在困境。

从垃圾信息积累端看，识别不足与动力欠缺造成了垃圾信息的持续堆积。一方面，无论是普通公众还是专业人士，均难以有效且准确地分辨 AI 生成的大量信息垃圾，进而导致低质量内容无法被有效分类与顺利清理。部分垃圾信息形式和内容与真实信息相似程度高，具有迷惑性，人们在识别过程中面临巨大挑战，传统的标注、审核与举报机制难以发挥全部作用。另一方面，平台对待垃圾信息的辨别与清理动力不足。鉴定并删除垃圾信息通常需要投入大量的人力、物力与技术资源，增加网络平台运营成本。同时，这些低质内容在一定程度上确实能够为平台带来流量的提升，在利益的驱动下，平台缺乏足够的动力去进行高效的分类和清理工作，从而导致垃圾信息不断积累，加剧信息环境的恶化。

2 AI 污染的法律治理现状与局限

针对 AI 污染及信息垃圾治理问题，相关规范散见于《数据安全法》《个人信息保护法》《著作权法》等法律及各地方性人工智能发展条例中。然而，设置这些规范的首要目的并非解决 AI 污染与垃圾信息传播问题，而是保障诸如数据权利、个人信息、商业秩序、知识产权等传统权利，规制 AI 污染仅是其附随效果。可见，现有模式缺乏规制直接性与体系性。

我国法律对 AI 污染的治理思路主要基于传统的损害填补与权利救济理念，对垃圾信息经传播后引起的损害后果加以调控。无论是民事、知识产权、数据侵权亦或是利用 AIGC 技术生成混淆内容扰乱市场正常竞争秩序等违法行为，法律介入时间普遍滞后于损害后果发生，存在明显滞后。此时，垃圾信息已被生成并留存于互联网中，并事实上持续污染互联网正常信息生态环境，即使主体责任严格落实，经济损失得以填补，由此形成的 AI 污染影响无法一并解决。由于网络数据与信息环境的庞大与复杂性，垃圾信息一旦出现于互联网中，就有被自动备份并收录进数据库的可能，即使于互联网中逐个删去相关侵权内容，仍无法排除下次生成内容不受该信息的影响。相反，对于表面上并不具有显著危害性的普通垃圾信息，其广泛生成、传播与堆积依然具有挤占数据存储、污染信息环境、增加用户阅读成本等负面影响，然而法律规范对其重视不足，更多依据平台与行业自律规定，通过系统标注或用户投诉、举报等方式加以治理，存在一定缺位。

从生成端与积累端看，尽管部分规范已指出平台方、研发者或使用者具有特定管理义务，但相关规范以部门规章和地方条例为主，缺乏专门立法。同时，现有规定多以倡导性宣言为主，对责任承担路径与具体后果缺少明确性规定，治理有效性不足。现有责任归属原则失灵构成数据污染治理的核心困境，难以平衡技术进步与社会安全 [2]。以《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》《上海市促进人工智能产业发展条例》等人工智能应用文件为例，就行为规范与应用方式，该类文件均指出各主体在应用 AIGC 等技术时应当遵守法律、法规规定，增强伦理意识，对生成内容负有监管与合规的相应义务，禁止实施危害法律法规和公序良俗的行为。但是对于具体法律责任承担形式，则规定笼统粗略。故这些文件虽然提供基本的治理框架，但部分条款设置与具体治理方式上存在不足 [3]。具言之，对于评估流程、监管流程与处罚措施等现实问题仍有赖于参考其他规范，其规定多为引致条款。常见表述为“依据《行政处罚法》和《刑法》”或“有关法律法规”予以处罚，缺乏细致落地指引。例如，《互联网信息服务深度合成管理规定》虽要求服务提供者与技术支持者对生成内容标注来源，但未明确标注具体标准及违法后果，实践中平台可能以“技术不可行”为由规避责任。[7]可见，相关规范从生成端与积累端切入调控的手段规范与治理力度不足，并导致污染治理路径终有赖于传统的事后权利救济模式。必须明确具体法律标准以实现风险应对与强化治理 [4]。

[7] 《互联网信息服务深度合成管理规定》第五条 服务提供者应当按照《互联网信息服务深度合成管理规定》第十六条的规定，在生成合成内容的文件元数据中添加隐式标识，隐式标识包含生成合成内容属性信息、服务提供者名称或编码、内容编号等制作要素信息。

《互联网信息服务深度合成管理规定》第二十二条 深度合成服务提供者和技术支持者违反本规定的，依照有关法律、行政法规的规定处罚；造成严重后果的，依法从重处罚。构成违反治安管理行为的，由公安机关依法给予治安管理处罚；构成犯罪的，依法追究刑事责任。

3 深化 AI 污染法律治理路径的讨论

如上所述，现有法律规制手段具有滞后性、间接性与低效性，有必要针对 AI 污染的特质进行深化治理。但是在讨论深化法律治理的具体模式前，我们应当认识到人工智能与 AIGC 技术的发展势不可挡，技术革新正在改变并重塑内容传播、消费习惯、认知理念等传统秩序，并为人类便利生活与社会高速发展带来切实益处。科学技术飞跃的过程中产生的负面影响与潜在风险本如“达摩克利斯之剑”，系科技发展的一体两面，无可避免。因此，如何让法律规范同时实现最大化抑制负面风险产生并保障科技自由发展的平衡目的值得讨论。

在 AIGC 技术诞生以前，网络垃圾信息就已广泛存在，只是 AI 技术发展使得“污染”以几何式速度扩张，公众辨别优质内容的难度与精力急剧提升，处理垃圾信息的能力已然不足。目前 AI 发展的最大困境在于用于 AI 学习的优质语料不足，而 AI 污染恰好加剧挖掘优质信息的难度 [5]。我们希望法律规范既能减少垃圾信息的生成、传播与积累，又不过分限制人类继续借助人工智能辅助创造内涵深刻的优质内容。为此，法律规制手段应当突破传统治理思路，适时介入调控 AI 污染不良影响与垃圾信息产生与堆积的全过程中。易言之，可以通过加强立法，深化司法，严格执法等法治手段方式，从 AI 污染生成端留痕、增强传播端监管、健全积累端同时处置，这将有助于降低治理成本，增强治理成效，促进治理模式体系化。

在生成端，应当积极立法，强化留痕要求。通过建立完善的生成端留痕系统使生成内容自动标注可溯源信息，为后续传播过程中的辨认提供明确标识。现有的《人工智能生成合成内容标识办法》已经对此进行框架性规定，但仍可进一步细化要求人工智能在技术允许的范围内，在其生成内容时附带特定的数字签名或专属代码，以记录生成时间、算法模型、数据源等关键信息，乃至制定具体、分级别的国家标识标准。这样不仅有助于追溯信息来源，还能提高公众对 AI 生成内容的警惕性和辨别能力。同时，留痕机制也为监管部门提供了有力的监管工具，便于对低质量生成内容进行快速溯源和有效查处，有效降低治理成本与难度。

在传播端，应当增强对不良传播的监管与违法打击，提升 AI 污染违法成本。一方面，应当在行政领域进一步明确平台对 AI 生成内容履行“分类审核义务”，如对新闻、医疗等高风险领域对于人工智能生成的信息加强技术标识，完善传播审查，健全版权规范，实施人工复核制度 [6]。对于故意、恶意大量传播 AI 污染信息人员，应于法律层面细化责任承担的具体形式与处罚标准，并严格落实相应责任，以形成有效威慑。同时，组建专业行政监管与执法团队，加强监管职责，落实行政处理流程与措施。另一方面，可以考虑对严重扰乱信息环境与数据安全的违法者加以刑事处罚予以严厉惩戒。尤其是对于部分利用 AI 技术恶意增加平台运营成本、骗取流量利益、污染数据库内容、严重误导公众认知等行为进行打击，以确保信息环境与社会秩序安全。可在刑事归责上可参照拒不履行信息网络安全管理义务罪，强化服务提供者的责任 [7]。

在积累端，平台和有关部门应共同努力，健全积累端处置措施，提升污染清理能力。通过扩充、完善垃圾信息投诉与举报形式与渠道，整合公众力量以强化垃圾信息清理工作。同时，增强平台的污染清理义务，落实企业责任，对低质量 AI 生成内容进行定期审核与及时删除。此外，可以成立由公众、专业人士与相关职能人员共同组成的专家小组定期评定网络 AI 污染情况与治理成效，为治理工作提供专业的评估和建议，并以其结论作为司法、行政执法中的关键证据。通过多主体参与的治理模式，提高治理的科学性和有效性。可参考韩国 2023 年成立的“AI 伦理委员会”模式，通过吸纳技术专家、法律人士与公众代表，制定行业标准并监督执行，明确将推动删除 AI 污染下的低质内容作为治理目标之一。[8]

4 结语

在 AI 技术迅猛发展的当下，AI 污染治理已成为亟待破解的关键命题。我们必须清醒地认识到，传统事后救济的思维难以有效应对 AI 污染的复杂性与动态性。应积极构建事前预警+积极监管的创新治理框架，以技术透明化作为降低治理成本的基石，借助法律条文的精细化精准界定各主体的权利与责任边界，同时充分激发

[8] 2024 年 4 月 7 日，韩国科学和信息通信技术部成立 AI 委员会以监督 AI 发展。该委员会将由部长李宗昊和大田大学校长南相豪主持，委员将从包括半导体行业在内的各个行业，以及金融、贸易、工业等其他政府部门中选出。该委员会拟对 AI 半导体、法律问题、伦理及安全问题进行研究，并重点关注并治理低质 AI 生成内容的“病毒式”传播问题，以确保人类从 AI 技术中获益。

社会共治的活力，凝聚政府、企业、科研机构、社会组织以及公众等多元主体的协同力量。从源头预防到过程监管，再到事后救济，打造“预防—监管—救济”全链条的闭环治理体系，以系统性思维推动 AI 污染治理的科学化、规范化与长效化。只有这样，才能在技术创新的浪潮中，切实维护公共利益，促进人工智能技术与社会发展的良性互动，实现人工智能产业的稳健、可持续发展，为人类社会的进步筑牢坚实根基。

参考文献

[1] 宋瑞,柳媛.警惕“AI 污染”乱象[N]. 新华每日电讯，2025-10-22.

[2] 曾庆醒.涌现效应下的生成式人工智能数据污染及其治理路径[J].法学杂志, 2024, 45(5):53-69.

[3] 宋华健. 论生成式人工智能的法律风险与治理路径[J].北京理工大学学报（社会科学版）, 2024, 26(3): 134-143.

[4] 郭建.论生成式人工智能的法律风险与治理路径[J].时代人物, 2024(8):0103-0105.

[5] 银昕.生成式 AI 乱相，AI“污染”当治[N].中国工业报，2024-11-11(01).

[6] 杜亚洁.AIGC 在新闻生产中的法律伦理与规制[J].法学, 2024, 12(11):5.

[7] 龚鹏程,傅成璐.生成式人工智能虚假信息的刑事风险与法律应对[J].江苏警官学院学报, 2024, 39(3):64-71.