公共产品理论下生成式人工智能虚假信息的层级治理研究

夏若曦

四川省社会科学院 四川 成都 610071

【摘　要】：生成式人工智能技术的快速发展在推动信息生产革新的同时也引发了虚假信息治理的复杂难题。当前治理困境源于生成式人工智能虚假信息的特殊性和复杂性，导致传统单一治理模式难以有效应对。公共产品理论为破解这一难题提供了新视角。通过将虚假信息依风险程度划分为纯公共产品层（高风险信息）、准公共产品层（中风险信息）和私人产品层（低风险信息），可构建分层治理框架。纯公共产品层需强化公权力主导，准公共产品层应压实平台技术核验义务，私人产品层需明确用户注意义务。

【关键词】：生成式人工智能；虚假信息；公共产品理论；分层治理

DOI:10.12417/3041-0630.26.06.083

1 问题的提出

生成式人工智能的爆发式发展，彻底改变了信息生产与传播的底层逻辑。这类技术依托大规模预训练模型与深度学习算法，能够快速生成文本、图像、音频等多模态内容，其产出效率与逼真程度远超传统人工制造模式。然而，技术红利背后潜藏着严峻的虚假信息风险。AI 生成内容可能包含逻辑合理但事实错误的表述，其虚假性与模糊性引发的风险已不容忽视。我国现行治理框架以 2023 年 7 月发布的《生成式人工智能服务管理暂行办法》为核心，该办法从内容生成、数据来源和主体责任等方面对虚假信息治理作出原则性规定。现行治理体系虽已初步构建多主体参与的治理框架，但在实践中暴露出显著的结构性矛盾。其一，高风险信息治理存在公权力介入不足问题。涉及公共安全的虚假信息，因跨平台传播特性与技术溯源难度，尚未形成全链条的法治闭环。生成式 AI 通过预训练模型与强化学习生成的虚假信息，其危害程度远超传统人工制造的内容，但实践中缺乏针对高风险信息的动态评估体系。其二，中风险信息治理面临平台技术核验义务虚化问题。平台对商业欺诈性信息的识别因缺乏明确的操作指引，导致规范效力难以落实。具体表现为：一是算法审核的技术局限性导致核验效能不足。二是实时监测的成本压力导致义务履行形式化。其三，低风险信息治理中用户注意义务边界模糊。现有法律尚未明确个人用户利用生成式人工智能的全过程注意义务。对于用户使用的故意诱导和过失误用责任承担规定不清晰，既未明确技术交互中诱导行为的具体表现形态，也未建立区分过失与故意的主观过错判定规则，导致实践中难以准确界定用户责任。

而公共产品理论的引入，为破解矛盾提供了新维度。根据该理论，社会产品可依消费非竞争性与受益非排他性程度，划分为纯公共产品、准公共产品与私人产品。将这一分析框架适用于虚假信息治理，可清晰界定不同风险层级信息的治理属性：高风险信息具有强公共危害性，属于纯公共产品范畴，需公权力以治理者角色统筹资源；中风险信息涉及特定领域秩序，为准公共产品，应由平台承担技术治理的主体责任；低风险信息主要侵害个体权益，属私人产品，需通过民事归责明确个体义务。

2 公共产品理论与虚假信息治理的逻辑勾连

2.1 公共产品理论的核心要义

公共产品理论的构建始于对社会资源分配本质的追问，其核心在于通过产品属性界定治理责任边界。萨缪尔森在《公共支出的纯理论》中确立的二分法，奠定了理论分析的基础框架。他提出，纯公共产品具有非竞争性与非排他性双重属性。前者指新增消费者不影响原有供给效率，后者指无法通过技术手段阻止未付费者使用；而私人产品则完全依赖市场定价机制，具备竞争性与排他性。这种二元划分揭示了政府与市场的基础分工。而后，布坎南的俱乐部产品理论进一步拓展了公共产品的谱系。他指出，现实中大量存在介于纯公共产品与私人产品之间的准公共产品。这类产品具有可排他性与有限非竞争性，即通过设定准入门槛实现排他，同时在消费容量上保持非竞争性。俱乐部产品的治理需引入多元机制。最后，奥斯特罗姆的多中心治理理论则突破了传统政府-市场的二元对立框架。他通过对公共资源治理的实证研究发现，纯公共产品的供给并非只能依赖公权力的单一中心治理，社群自治和非营利组织参与等多元主体协同模式同样有效。这一理论强调，治理效能的提升需打破公权力万能的假设，构建政府、市场和社会相互补充的治理网络，尤其在信息不对称或技术复杂的领域，多元主体的知识互补性至关重要。

作者简介：夏若曦，（1999.11—）女，汉族，四川成都人，硕士研究生，四川省社会科学院，研究方向：民商法。

而公共产品理论的引入为生成式人工智能虚假信息治理提供了全新的分析范式。通过经典理论内核与数字治理场景的双重标准，可清晰界定不同风险层级信息的治理属性，避免一刀切式监管对技术创新的抑制，同时通过分层治理实现资源最优配置，最终在技术发展与风险防控之间建立动态平衡。

2.2 虚假信息的公共产品属性分型

公共产品理论对社会产品的分类逻辑，可用于解析生成式人工智能环境下虚假信息的治理属性。根据虚假信息的传播范围、危害程度及治理主体的责任差异，可将其划分为以下三个层级。其一是纯公共产品层，即高风险信息的全域性危害。纯公共产品层的虚假信息具有完全非竞争性与非排他性特征，其传播直接威胁国家安全、公共安全或社会基本秩序。此类信息的危害后果无法通过个体或市场机制自行消化，治理收益覆盖全体社会成员且不可分割。从治理责任看，该层级信息属于典型的治理公共产品，需以公权力主导全域治理。通过立法、行政监管与技术标准制定，对高风险信息的生成源头实施严格管控。同时应当建立全流程监管机制，从数据训练、算法设计到内容输出实施穿透式审查，确保此类信息的治理具有强制力与普遍约束力。其二是准公共产品层，即中风险信息的领域性影响。准公共产品层的虚假信息具有有限非竞争性与可排他性特征，主要影响特定领域的秩序或特定群体的利益。此类信息的传播范围通常限定于特定平台或群体，但其危害可能通过跨平台扩散转化为公共风险。该层级信息的治理需依托公权力和平台协同机制，其中平台应当尽到更为扎实的技术核验义务与算法透明度义务。可要求平台运用生成式人工智能技术对商业广告、新闻资讯等内容进行实时核验，通过多模态特征比对、传播路径分析等手段识别异常信息。其三是私人产品层，即低风险信息的个体性侵害。私人产品层的虚假信息主要体现为竞争性与排他性特征，其传播以个体权益侵害为导向。此类信息的危害通常限于特定主体，治理收益可通过个体维权实现排他性享有。该层级信息的治理责任主体为用户个体，法律需明确其注意义务与侵权责任边界。可通过界定个体用户的主观过错标准，判断用户对低风险信息是否存在主观过错。用户利用生成式工具制作虚假信息，若技术手段足以识别信息伪造痕迹，用户需承担故意侵权责任。

3 公共产品视域下生成式人工智能虚假信息分层治理的规制路径

3.1 纯公共产品层：公权力主导

首先，建立虚假信息的等级评估制度。我国现行立法已在《互联网信息服务算法推荐管理规定》第 27 条、《互联网信息服务深度合成管理规定》第 20 条和《生成式人工智能服务管理暂行办法》第 17 条均要求具有舆论属性或者社会动员能力的技术服务应当开展安全评估。但现有评估聚焦于服务内容审查与风险防控能力，缺乏对虚假信息本身的层级化评估。基于公共产品理论对纯公共产品的属性界定，高风险信息的治理需构建等级评估体系。参考欧洲议会《人工智能法案》的风险分级逻辑，根据侵害利益的严重程度可将生成式 AI 生成的虚假信息划分为三级：高等级信息涉及直接危害国家安全和社会稳定的内容，其传播将引发系统性风险，属于纯公共产品层治理核心对象，应当从优从重处罚；中等级信息涉及损害社会公共利益的内容，具有准公共产品属性，需平台与公权力协同治理；低等级信息涉及侵害个体权益的内容，属于私人产品层，以个体追责为主。

其次，建立特殊虚假信息的联合评估制度。现有立法对涉及生物识别信息和国家形象的特殊信息，仅要求服务提供者自行或委托评估，但自我定义评估模式难以保证中立性。特殊信息评估需实现三要求：一是评估主体多元化。参照 2021 年《关于加强互联网信息服务算法综合治理的指导意见》，组建由技术专家、行业组织以及第三方机构构成的联合评估组，避免单一主体评估的局限性。二是评估方式标准化。制定《特殊虚假信息评估技术规范》，明确生物识别信息伪造和国家形象损害等场景的评估流程。同时，评估需覆盖数据训练、模型优化和内容生成全链条，而非仅针对最终输出；三是合成数据专项评估。针对生成式 AI 大量使用合成数据的特性，建立独立评估规则。合成数据因融合原始数据、中间数据与输出数据，其虚假信息风险具有累积性，需评估合成数据的来源合法性、训练过程偏差率以及与真实数据的混杂比例。评估结果需实行双报告制：技术评估报告由第三方机构出具，法律风险报告由司法机关参与制定，确保技术合规性与法律合法性的双重审查。

最后，强化前置性防控措施。现行规定对高风险信息的规制侧重事后处置，缺乏源头阻断机制。因此，根据人工智能运行的特殊性，前置防控需从两个阶段介入：一是模型训练阶段：要求服务提供者在预训练数据中剔除高风险信息源，对涉及国家敏感领域的数据实施负面名单管理；二是模型部署阶段：建立高风险信息生成熔断机制。参考欧盟《人工智能法》对不可接受风险技术的禁止性规定，要求服务提供者在模型中嵌入实时阻断模块，当检测到高等级信息生成指令时，自动终止运算并上报监管部门。

3.2 准公共产品层：平台守门

在准公共产品层（中风险信息）的治理中，针对平台技术核验义务虚化的现实困境，需以公共产品理论为基础完善技术标准和平衡核验成本，以此达到实化平台核验义务的效果。平台作为数字居间方应当承担虚假信息的技术核验责任，需通过法律规制促使平台在中风险信息治理中发挥技术枢纽作用，弥补公权力监管与个体自治之间的治理断层。

其一，完善动态更新的技术标准体系。针对算法审核技术局限性的问题，需完善动态更新的技术标准体系。可根据算法透明度义务的内在需求具化技术标准的核心要求，即要求平台采用多模态特征融合技术，并定期向监管部门提交技术效能报告。现行《生成式人工智能服务管理暂行办法》第 12 条虽要求内容标识，但未明确技术实现路径，可在此基础上增设中风险信息技术核验规范，强制平台采用基于深度学习的语义理解模型，提升对语义变形虚假信息的识别能力。同时可以现有技术水平为标准来判断核验技术是否合规，即平台需证明其技术措施达到行业同期平均水平，否则需承担过错推定责任。

其二，构建定向补贴与市场激励机制。为解决实时监测的成本压力问题，需构建定向补贴与市场激励的成本分担机制。一方面，通过专项基金对平台采用的高精度检测技术给予研发补贴以降低企业技术投入成本；另一方面，允许平台将合规成本纳入服务定价，通过《暂行办法》第 17 条的安全评估机制，对达到高核验标准的平台给予税收优惠或市场准入便利。同时平台若能证明其在合成数据识别技术上的投入，可相应减轻对该类信息的核验责任，形成激励相容的治理环境。

3.3 私人产品层：用户自律

一方面，细化用户注意义务的内容与判定标准。用户注意义务的具体化需围绕技术交互的全流程展开。在信息输入环节，应明确用户负有合理克制义务，即不得为获取特定虚假信息而实施具有明显诱导性的输入行为。此处的诱导性判定可从主客观双重标准构建：主观上以用户是否明知或应知其输入内容可能导致虚假信息生成为要件，客观上以输入行为是否超出正常信息查询的合理边界为标准；在信息传播环节，需建立二次传播注意义务，要求用户对生成内容进行必要的真实性核验，尤其当内容涉及他人人格权益或重大公共利益时，用户未经核实的主动传播行为应被认定为违反注意义务。此外，针对低风险信息的特性，可引入行业自律标准作为辅助判定依据，通过生成式人工智能服务提供者制定的用户协议、使用指南等，细化不同场景下的注意义务内容。

另一方面，构建用户行为的分层责任体系。用户责任的认定需区分行为性质与危害后果，建立阶梯式责任架构。用户利用生成式人工智能生成或传播虚假信息的行为，本质上是个体借助技术工具实施的民事侵权活动。若用户在交互中故意采用诱导性输入，或明知生成信息存在虚假性仍主动传播，其主观过错与行为后果已符合一般侵权行为的构成要件。此类故意诱导生成虚假信息并造成他人权益损害的行为，应依据《民法典》第 1165 条等规定认定其构成一般侵权责任，用户需承担停止侵害和赔偿损失等民事责任；若行为涉及行政违法，则结合《网络安全法》相关规定追究其行政责任。在主观过错认定上，区分故意诱导与过失误用：前者指用户明知输入行为会导致虚假信息生成仍积极追求结果，后者指用户因疏忽大意未履行合理核验义务，并据此设置不同的责任承担比例。

4 结语

生成式人工智能技术的迭代发展在重塑信息生产范式的同时，也使虚假信息治理陷入传统规制模式的失灵困境。本研究基于公共产品理论的分析框架，将生成式 AI 虚假信息依风险属性划分为纯公共产品、准公共产品与私人产品三层，构建了公权力主导、平台核验和个体担责的分层治理体系。这一模式构建的核心价值在于，突破了传统一刀切监管的制度局限，通过治理责任的类型化分配实现资源优化配置。而从规范实践看，分层治理框架的落地需依托规范体系的协同完善：纯公共产品层需强化专业机构风险评估与前置防控的制度刚性；准公共产品层需通过技术标准与激励机制实化平台核验义务；私人产品层则需以行为规范与责任划分明晰用户自律边界。这种基于产品属性的分层规制，既回应了生成式 AI 技术的复杂性特征，也为数字时代信息治理提供了新型参考范式。

参考文献：

[1] 张素华,李凯.生成式人工智能虚假信息风险与治理研究[J].学术探索,2024,(07):129-140.

[2] 吴昺兵,贾康.政府与市场合作供给公共产品的理论分析和制度设计[J].江西社会科学,2023,43(05):157-171.

[3] 张乾友.解构公共性:对几种流行观念的批判性分析[J].学习论坛,2023,(02):58-64.

[4] 王利明.生成式人工智能侵权的法律应对[J].中国应用法学,2023,(05):27-38.

[5] 胡泳.人工智能驱动的虚假信息:现在与未来[J].南京社会科学,2024,(01):96-109.

[6] 张文祥.生成式人工智能虚假信息的舆论生态挑战与治理进路[J].山东大学学报(哲学社会科学版),2025,(01):155-164.