生成式 AI 的伦理态度研究：基于隐私，算法透明度与安全的多维度分析

唐浩瀚

大连理工大学，辽宁省大连市，116000；

摘要：[研究目的]探讨生成式人工智能在新闻传媒领域的应用及其引发的伦理挑战，分析其在提升新闻生产效率的同时如何通过隐私保护、算法透明度和内容安全治理，平衡技术创新与社会伦理风险，以构建安全、公平、透明的人机协同生态。[研究方法]采用理论推演与实证分析相结合的方法。通过文献梳理与政策文本解析,构建 AIGC 伦理治理的理论框架；结合典型案例实证分析技术应用中的隐私泄露、算法偏见及新闻伦理失序问题。[研究结论]AIGC 的健康发展需以“技术-制度-社会”协同治理为核心：强化算法可解释性工具与数据血缘追踪系统，建立分级监管框架；推动跨学科伦理委员会建设，完善公众数字素养教育；最终通过伦理约束与技术创新双轮驱动，在保障新闻真实性与公信力的前提下释放媒介新质生产力。

关键词：生成式人工智能；新闻伦理；算法透明度；隐私保护；人机协同；技术治理

DOI：10.69979/3041-0673.25.03.004

自 2022 年 11 月以来，以 ChatGPT 为代表的生成式 AI（Generative AI）逐渐走进人民的视野并深刻的参与和影响社会民生的各个领域。中华人民共和国工业和信息化部门联合其他多个部门于 2023 年发布的《生成式人工智能服务管理暂行办法》提出，“鼓励生成式人工智能技术在各行业、各领域的创新应用，生成积极健康、向上向善的优质内容，探索优化应用场景，构建应用生态体系。支持行业组织、企业、教育和科研机构、公共文化机构、有关专业机构等在生成式人工智能技术创新、数据资源建设、转化应用、风险防范等方面开展协作”。

虽然人工智能早已在专业的信息生产领域开始使用和探索，如军工领域的国产 GL6 主动防御系统通过人工智能判断炮弹或无人机预定轨道实现 180 度拦截，但成为普通人的信息生产辅助工具却是近两年的事情。非专业的普通信息生产者对于生成式 AI 的实现原理不明确，使得通过 AI 得到的人工智能生成内容（Artificial Intelligence Generated Content，AIGC）对于普通人来说更像是技术黑箱：生成内容由内在的算法与数据控制，人工智能算法在自主处置海量数据信息的过程中，依据复杂的运行逻辑形构了不确定的归纳知识，使得人们难以明确其中的因果关联，难以考量算法系统内部的调节方式，难以辨明算法结果或决策应用失误的本末源流。[1]

尽管人工智能大语言模型（Large Language Models，LLMs）处于快速迭代与规模化应用阶段，但其训练数据在文化覆盖广度与知识更新时效性维度仍存在显著局限：实时动态信息的捕获与整合机制尚未形成完备的技术闭环。同时，AIGC（人工智能生成内容）系统在数据安全性与算法科学性层面存在结构性缺陷，易诱发生成内容失真（即"数字幻觉"）及技术滥用风险（如 Deepfake 技术的非法应用）。更值得注意的是，社会公众普遍存在技术认知滞后性与批判性思维缺位现象，多重矛盾的交织将加剧技术伦理框架失序，最终导致数字时代的"技术异化风险"与社会信任危机。

1 AIGC 面临的伦理困境

AIGC 被认为是继专业生成内容（Professional Generated Content，PGC）和用户 生成内容（User Generated Content，UGC）之后的新型内容创作模式。[2] 在此过程中，技术应用需超越基础性的功能适配（如效率提升与流程优化），更需构建涵盖数据隐私保护、算法可解释性、偏见消减与社会价值校准的多维治理框架。这要求政策制定者与技术管理者深度理解人工智能的技术逻辑与利益相关者诉求，通过预研伦理风险、建立弹性监管机制，确保技术发展遵循安全、公平、透明及人本主义原则，从而规避潜在的算法歧视、信息操纵等社会危害。英国防务智库皇家联合军种国防研究所（RU

SI）指出人工智能的政策均须考虑必要性、合理性、透明度、问责制及其入侵性。

1.1 数据合规性缺失与隐私权侵害风险

人工智能算法的智能性表征本质上是数据驱动范式的具象化体现——算法模型的泛化能力与预测精度高度依赖于训练数据的规模性、多样性与时效性。在技术实现路径上，数据不仅构成算法训练的初始参数空间（parameter space），更是驱动算法通过自监督学习（self-supervised learning）实现持续优化的核心生产要素。这一特性直接催生了算法研发者的"数据饥渴"（data hunger）现象：为突破模型性能的帕累托边界（Pareto frontier），开发者往往采取大规模数据爬取、多源异构数据融合等策略以扩充训练集，典型案例包括 Google 的 BERT 模型通过吞噬超 33 亿词汇量的语料库实现语义理解突破，以及 OpenAI 的 GPT-4 依托 45 TB 多模态数据进行预训练。当前，算法所依赖的数据大部分来自传感器和人，包括用户上网新闻浏览记录、社交网络记录、网络购物记录、通讯软件聊天记录、传感器数据和监视数据等，[3]其中不可避免地涉及用户隐私信息。就算法而言，它是“数据贪婪”的（data-greedy），它驱动了算法研发机构和算法使用机构去无限制地收集、处理、使用个人信息。

算法驱动下的大规模数据收集带来的挑战表现在两个方面：一是数据受托方可信性缺失导致的多级泄露危机。当数据收集主体存在安全管控失效时，原始数据可能通过供应链漏洞发生级联扩散，形成“个人-组织-国家”三层风险传导，例如 2023 年 Telegram 黑客利用外卖平台数据接口漏洞，窃取 920 万用户配送地址与消费偏好数据共计 45 亿条地址信息泄露，直接导致精准诈骗案件激增；二是模型可逆性引发的敏感信息萃取风险。攻击者可借助模型反演攻击（Model Inversion Attacks）与梯度泄露攻击（Gradient Leakage Attacks），从黑盒化算法输出中重构训练数据特征。因此，有观点认为，个人数据和隐私保护不仅应该借助法律法规的约束，也应该在算法的设计、训练和部署过程中保证个人隐私数据不被未授权的人员直接或间接获取。[4]

对于受众来说，对于 AI 有着相同的担忧。一份由清华大学大数据治理研究中心进行的调查显示，公众对算法应用中的个人信息保护普遍存在较高焦虑，尤其对隐私泄露和信息盗用两大风险最为敏感，超过 80%的受访者明确表达了对这两类问题的担忧。

1.2 算法黑箱效应与歧视性决策危机

算法黑盒效应形成源于多重因素交织作用：首先，人类认知能力与复杂算法系统之间存在解析鸿沟，即使采用可解释性工具也难以完全追溯决策逻辑；其次，自学习算法的动态特性导致决策规则持续演变，开发者难以实时追踪其逻辑变更；最后，算法的持续迭代特性造成版本历史断层，使得用户无法回溯算法演化路径。这三重机制共同构成了技术透明化治理的核心障碍。

算法的不透明的直接结果就是可解释性更加低下。2017 年 ICML 的 Tutorial 中给出的一个关于可解释性的定义是：“解释是给人类作出解释的过程”引申来说就是人类能理解的描述给出解释，以让人类能看懂。算法之所以难以解释，是因为“黑箱”现象的存在。机器学习基本就是线性数学，很好解释，但是一旦涉及多层神经网络，问题就变成了非线性数学，不同变量之间的关系就纠缠不清了。

生成式人工智能的涌现能力（Emergent Capability）与自迭代机制（Self-Optimization）导致传统伦理治理框架失效——其决策过程既无法通过预设行为准则实现价值对齐（Value Alignment），也难以通过道德代码嵌入保障输出结果的公平性。这种技术特性引发的社会伦理困境在多个应用场景中显性化：例如，文本生成模型可能系统性强化性别职业偏见，例如将“护士”关联为女性形象，图像合成算法倾向于将高收入社区表征为白人主导的居住空间。由于生成式 AI 的决策黑箱特性，例如 Transformer 架构中多头注意力机制的不可解释性，受其影响的用户往往无法追溯歧视性结果的生成逻辑，包括训练数据的偏见渗透路径、上下文感知的权重分配规则及潜在社会刻板印象的强化机制。根据欧盟《人工智能法案》，开发者需向用户提供可理解的决策解释，涵盖模型的功能边界、数据偏见校正方法及输出结果的归因分析，否则将触发算法歧视的“合理性危机”，加剧社会信任体系的结构性瓦解。

简言之，算法可解释性的目的包括维护算法消费者的知情权利益，避免和解决算法决策的错误性和歧视性，明晰算法决策的主体性、因果性或相关性，进而 助力解决算法可问责性问题。

虽然算法透明的呼声很高，但是有学者认为算法透明绝不是一个道德准则，而是一个“有利于道德条件”；

相反，绝对的透明本身就会形成一个道德问题。[5]因为算法本身也存在安全需求：在商业竞争维度，透明的算法允许第三方主体能够规避数据采集与模型训练的前期投入，通过 API 接口劫持等技术剽窃手段实施低成本竞争，从而侵蚀算法所有者的市场独占性收益；在隐私安全视角，算法训练集常包含用户敏感个人信息，一旦因数据脱敏失效或系统漏洞导致信息外泄，将显著提升网络诈骗、精准勒索等犯罪行为的实施概率。因此，报告显示，算法设计过程，尤其是源代码，多数人并不认为应该公开。

1.3 新闻伦理失序与媒体信任崩解

生成式人工智能（AIGC）的技术特性正在重塑数字内容生产与传播的伦理边界。相较于传统 AI 的透明度与隐私争议，以 Stable Diffusion 为代表的制图 AI 生成式技术催生了更具破坏性的"深度造假危机"——其核心矛盾在于超现实内容生成能力与事实核查机制间的结构性失衡。通过多模态对齐（Multimodal Alignment）技术，AIGC 可批量生成语义连贯的虚假政治宣言、物理精确的灾难现场图像，以及情感共振的深度伪造视频。斯坦福大学计算政策中心 2023 年研究显示，AIGC 生成的虚假新闻在社交媒体平台的传播速度较人工编造内容快 17 倍，且被举报概率降低 62%，印证了技术赋能的"谎言通胀"（Liar’s Dividend）效应。

2 AIGC 伦理困境的治理策略

目前，人工智能正在深刻影响着人类社会的各个领域。它不仅推动了社会经济的发展，提高了生活质量，同时也带来了一些伦理问题，涉及安全性、透明度、公平性、隐私保护以及人类尊严等方面。国际社会普遍认为，人工智能的发展亟需伦理价值的引导与约束，世界各国也已采取了多种措施应对由人工智能带来的伦理挑战。

2.1 构建人工智能伦理治理架构：强化伦理委员会职能与基本原则建设

人工智能的伦理风险在产品研发与应用的各个环节中广泛存在，问题的复杂性和涉及的价值判断高度交织。当前，研发人员普遍缺乏系统的伦理知识，难以有效承担关键伦理决策的责任。在此背景下，成立专门的人工智能治理机构，尤其是在政府和企业层面，已成为推动人工智能治理体系制度化与专业化的关键途径。

当前，相关机构已经逐渐在国家、企业与社会团体等组织层次成立。法国于 2019 年 4 月组建了“人工智能伦理委员会”，旨在监督军用人工智能的发展。而在中国，国家科技伦理委员会已于 2019 年 7 月 24 日宣布组建，旨在对包括人工智能在内的一系列科技伦理问题展开制度化治理。在企业治理层面，建立人工智能伦理治理架构已成为全球科技行业履行技术伦理责任的核心机制。以 Meta 的负责任创新中心为例，其通过跨部门协作审查算法偏见与内容治理风险；亚马逊则设立人工智能伦理委员会，对云服务客户端的 AI 应用进行合规性审计。此类机制通过制度化设计将伦理原则嵌入研发全流程，形成行业自律的基础框架。中国科技企业亦加速构建本土化治理体系：百度于 2023 年成立科技伦理委员会，由首席技术官领衔并引入法学、伦理学领域专家顾问；腾讯组建的 AI 治理实验室不仅制定内部《人工智能应用准则》，还牵头参与 IEEE 全球伦理标准制定工作。这些实践表明，组织化、专业化的伦理治理架构正在成为企业应对技术不确定性的关键基础设施。在社会团体层面，中国智能科学技术领域唯一的国家级学会——中国人工智能学会（CAAI）于 2018 年年中组建了人工智能伦理专委会。

2.2 完善隐私保护制度体系：健全法律法规与监管政策执行机制

传统法律体系的规范框架在个人信息保护维度呈现出明确的层级化特征：从法理层面分析，其保护客体主要聚焦于私人领域内具有身份识别效力的信息载体，并依据信息敏感度差异构建梯度化规制模式。规范架构层面，法律文本通过引入"普通个人信息"与"敏感个人信息"的二元分类体系，建立起差异化保护机制。对于涉及种族、生物特征、健康数据等敏感信息类别，立法确立了更严格的处理要件——需以明示同意为前置条件，或在重大公共利益等法定豁免情形下方可实施处理行为。技术管控层面，法律强制要求数据处理者对敏感信息实施加密存储机制，并构建多因子身份验证系统以强化访问权限控制。值得注意的是，现行规范体系以"可识别性"作为法律保护的核心要件，当数据经过去标识化技术处理，如泛化算法、差分隐私、合成数据生成等技术路径，其法律属性将发生根本转变——此类技术处理使数据主体无法被直接或间接识别，因而脱离个人信息保护法的规制范畴，相关数据的流通利用即进入法

律许可的开放空间。

2017 年 12 月 29 日，全国信息安全标准化技术委员会主导研制的《信息安全技术 个人信息安全规范》经国家标准化管理委员会批准发布，该标准于次年 5 月 1 日正式生效施行。作为我国首部系统性规制个人信息处理活动的技术标准，其突破性贡献在于通过标准化的技术路径，首次在规范性文件中明确定义"个人敏感信息"的构成要件与识别标准，填补了个人信息分类保护制度的空白。该标准在司法实践中被赋予准规范性效力，不仅为《网络安全法》配套实施细则的制定提供技术支撑，更成为后续《个人信息保护法》立法的重要法源基础，标志着我国个人信息保护体系从分散式立法向技术标准与法律规范协同治理的范式转型。

2.3 深化算法技术研发创新：推动可解释性与透明性技术突破

技术产出的问题，固然需要法律规制、伦理规制甚至意识形态规制，但技术产生的问题还是需要技术本身来解决，正如马克思所言："批判的武器当然不能代替武器的批判，物质力量只能用物质力量来摧毁。"[6]有学者提出"算法向善"，让社会力量主导算法的发展方向，让算法更好地服务人类社会，回到以人为本的向善。[7]提高算法的透明度，增加算法的可解释性，例如公开算法决策体系。[8]

另一种较新也较有潜力的增强算法透明度的方法是，使用技术工具来测试和审计算法保持高度的透明度。建立数据血缘追踪系统（Data Provenance Tracking），如 IBM 的 AI FactSheets 记录训练数据来源与特征工程过程，可将数据偏见识别效率提高 63%。欧盟《数字服务法案》要求平台算法需保留 6 个月内的决策日志供监管审计。

此外，在算法日益渗透日常生活的情况下，有必要加强公共教育，以提高普通公众计算素养和数据素养，以更好地实现算法的可解释性。比如，纽约大学的 AI Now 研究所制定了算法影响评估指南，旨在提高公众对机器学习算法的认识和讨论。[9]

3 总结

生成式人工智能作为技术工具，正在深刻重构新闻生产的全流程，推动人机协同模式下媒介新质生产力的释放。通过自动化采编、智能分发与多模态内容生成，显著提升了新闻生产效率，但技术赋能的背后潜藏着不容忽视的数字伦理风险。未来，AIGC 的健康发展需在效率与伦理间寻求动态平衡——既不能因技术风险否定其革新潜力，亦不可为追逐流量放任"合成内容泛滥"。唯有坚持"以人为本"的技术伦理观，通过技术创新与制度约束的双重驱动，方能在智能时代守护新闻真实的生命线，重塑更具韧性与公信力的媒介生态。

参考文献

[1]黄静秋,邓伯军.人工智能算法的伦理规制研究[J].北京科技大学学报(社会科学版),2025,41(2):88-96.
[2]孙玉明,张子怡.智媒时代生成式 AI 的伦理风险及其治理路径[J].新闻论坛,2024,38(06):26-29.
[3]刘雅辉,张铁赢,靳小龙,程学旗.大数据时代的个人隐私保护[J].计算机研究与发展,2015,52(01):229-247.
[4]刘俊旭,孟小峰.机器学习的隐私保护研究综述[J].计算机研究与发展,2020,57(02):346-362.
[5]Turilli M,Floridi L.The ethics of information transparency[J].Ethics and Information Technology,2009,11(2):105-112
[6]马克思，恩格斯.马克思恩格斯全集（第 3 卷）[M].北京：人民出版社，2002.
[7]邱泽奇.算法向善选择背后的权衡与博弈[J].人民论坛,2021,(Z1):16-19.
[8]丁晓东.论算法的法律规制[J].中国社会科学,2020,(12):138-159+203.
[9]Tsamados A, Aggarwal N,Cowls J,et al.The ethics of algorithms:key problems and solutions[J]. AI &SOCIETY, 2021:1-16

作者简介：唐浩瀚（2003-）男，汉，江苏宿迁，本科，研究方向：数字媒体技术