基于算法机理的生成式人工智能伦理评价指标体系构建与治理研究

陶婷婷　李本乾

【摘　要】生成式人工智能风险正引发日益广泛的伦理关切。当前的相关研究多集中于针对伦理现象的描述，鲜少剖析其背后的深层次技术逻辑与风险成因，导致对伦理风险缺乏全局性的系统研究。研究首先提出了基于算法机理和数据驱动的生成式人工智能伦理指标体系构建方法，提供了新的伦理研究视角和方法论；其次，研究构建了一个涵盖多层级的生成式人工智能伦理指标体系，清晰勾勒了生成式人工智能的伦理风险全景，并明确了治理的重点方向；最后，研究在技术与社会层面提出了针对性的伦理风险治理路径，为相关政策的制定提供参考。

【关键词】生成式人工智能；伦理评价体系；伦理评价指标；伦理风险
【中图分类号】G21　　【文献标识码】A

一、引言

生成式人工智能的崛起标志着人类朝着通用型人工智能（AGI）跨出了一大步，而其带来的潜在道德伦理问题也正引发越来越多的伦理关切。正如OpenAI所言：“作为机器学习模型，ChatGPT没有理解或考虑伦理和法律问题的能力。它无法根据伦理或法律原则做出判断或决策”。[①]以ChatGPT为代表的生成式人工智能带来的威胁包括但不限于：算法黑箱、隐私安全、歧视和偏见、粗俗语言、版权侵权、抄袭、欺诈交易、虚假信息等[②]，处理其带来的伦理问题和潜在陷阱变得至关重要。

当前的伦理研究多集中于对生成式人工智能伦理问题的现象描述，缺乏对其背后深层次技术逻辑与风险成因的深入分析，导致研究的全局性和系统性不足。本文在剖析生成式人工智能技术机理的基础上，构建多层级生成式人工智能伦理指标体系，不仅有助于保障人工智能生成过程符合伦理规范，也有效规避了人工立场的偏见，确保了指标体系构建的科学性和严谨性，为生成式人工智能的伦理治理提出有针对性的框架。

二、理论基础

在生成式人工智能兴起之前，人工智能伦理问题研究早已全面铺开。《G20人工智能原则》和欧盟《人工智能法案》[③]等先后提出了人工智能伦理指南和原则。Müller[④]区分了人工智能系统作为对象的伦理问题，包括隐私、操控、不透明性、偏见、人机交互、就业和自主性影响，以及人工智能系统作为主体的伦理问题，如机器伦理、人工道德代理等。Vesnic-Alujevic等人[⑤]进一步划分了伦理问题的范畴，包括个人层面的自主性、尊严、隐私与数据保护，以及社会层面的公平与公正、美好生活与多样性、责任与问责、透明度、监控与数据化、人工智能治理等。

随着生成式人工智能的发展，伦理问题讨论进一步延伸到大语言模型等生成技术带来的伦理议题。例如，经济合作与发展组织（OECD）《人工智能语言模型》在原有的原则基准上，提出了与语言模型等生成式人工智能相关的伦理考虑因素；Shmueli等[6]探讨了在学术出版领域应用ChatGPT等生成式人工智能产生的伦理挑战，涵盖了作者身份、准确性、隐私、操纵系统、问责等；Susarla等[7]深入分析了生成式人工智能面临的挑战，其中涉及训练数据中的偏见和幻觉、真实性和可靠性、知识产权，以及内容深度性不足等问题；Dwivedi等则指出ChatGPT涉及隐私、偏见、透明度、治理、劳工影响等重要伦理问题；[8]谷歌旗下的人工智能公司DeepMind也在其报告[9]中提出了21种可以从语言模型中预期到的伤害风险，涉及歧视、排他和有毒言论；信息泄露危害；错误信息危害；恶意使用；人机交互危害；环境和社会经济危害六个风险领域。[10]

以上生成式人工智能伦理问题的研究提供了一定理论基础，但整体上缺乏系统性的方法和框架。例如，尽管DeepMind的报告提出了语言模型的风险分类，但并没有为如何系统地识别这些风险提供明确的方法和指引。若要有效引导和监管生成式人工智能技术的发展，亟需建立一套科学系统的伦理评价指标体系，以规范技术的应用，降低潜在的伦理风险。

康德伦理学为构建这一伦理指标体系提供了重要的启示。作为规范伦理学中义务论的典型代表，康德为程序公平（Procedural Fairness）提供了哲学基础。康德视角认为，应该有一套程序性原则或规则（A set of procedural principles or rules），以系统地指导人们做出符合道德的行为。例如，尽管入狱是一个不理想的结果，但如果是由律师和法官正确遵循程序做出的决定，那么此决定通常被认为是公平的。[11]

程序公平不仅要求结果本身的正当性，更强调实现这一结果的过程是否符合伦理标准。基于这一理念，本文将在深入分析生成式人工智能算法机理及其可能引发的伦理问题的基础上，构建一套涵盖生成式人工智能整体程序流程的伦理指标体系，确保在数据收集、模型训练、结果生成等环节中，系统能够避免偏见和歧视等伦理风险，实现程序上的公平与正义，因为“如果整个流程符合特定的伦理要求，那么设计符合程序公平的人工智能系统是可能的”。[12]

三、研究思路

既往关于人工智能伦理的研究常使用德尔菲法或情景研究（Scenario Research）。[13][14]其中德尔菲法作为一种社会统计研究方法，常用于构建社会领域的伦理指标体系。然而，由于该方法依赖于专家的主观判断，并且受限于专家数量，难以全面覆盖所有潜在的伦理风险。因此，有必要探索更加系统的方法，以更加客观全面地识别和评估生成式人工智能的伦理风险。

本文的体系构建旨在解决以下三个方面的问题：（1）如何获得科学合理的生成式人工智能伦理评估指标？（2）这些评估指标是否具有普适性？（3）在前两个问题的基础上，如何为指标赋予恰当的权重，构建完整的评估指标体系？笔者认为，一个科学合理的伦理指标体系应当覆盖多个维度，并满足以下三项基本原则：（1）程序正义：内容生成的算法流程应符合伦理要求，因为只有实现这一点，才可能设计出“符合程序公平的人工智能系统”；[15]（2）全面性：伦理指标应系统全面，避免遗漏；（3）科学性：应通过科学系统的方法确保指标体系的严谨。

基于以上原则，研究设计包括以下步骤：（1）初步构建：基于算法程序识别潜在伦理风险，初步建立伦理指标；（2）指标完善及体系建构：结合

当前研究成果，通过数据驱动的主题分析（Topic Analysis）和聚类方法，补充、建构伦理指标体系；（3）相关性检验：将本文构建的伦理指标体系与语料文献的伦理主题进行相关性对比，验证指标体系的科学有效性；（4）指标赋权：通过归一化（Normalization）方法计算指标权重，完成评估指标体系的赋权。通过这种方法构建的生成式人工智能伦理指标体系，尽可能减少了德尔菲法等专家调查法中的主观影响，确保了科学性、客观性和全面性。

四、生成式人工智能伦理评价指标体系构建

（一）初步构建

威廉·盖弗（William W. Gaver）在其“技术可供性”观点中指出，“可供性让我们不再只关注技术或只关注使用者，而是关注两者之间的互动。”[16][17]这意味着技术不仅提供了某种功能或工具，也通过与人类的互动塑造着各领域潜在的伦理风险。这一观点为理解新兴技术的伦理问题提供了重要的理论基础。正是在这种对技术与人类互动的关注下，近年来国际上涌现了一系列“新兴技术的伦理研究方法”（Methodologies for Studying the Ethics of Emerging Technologies）[18]，包括“预见性技术伦理”（Anticipatory Technology Ethics，ATE）[19]、“信息技术伦理影响评估框架”（Ethical Impact Assessment，EIA）[20]和“新兴ICT应用的伦理问题”（Ethical Issues of Emerging ICT Applications，ETICA）[21]等。这些方法对生成式人工智能伦理研究提供了重要的理论支持。ATE从技术伦理的哲学角度出发，涵盖了危害和风险、权利、（分配）正义、福祉与公共利益等主要类别；EIA包括自主权、不伤害、行善和正义，并补充了隐私和数据保护等原则；ETICA则基于对十种不同新兴技术的分析，列出了概念问题和伦理理论。这些伦理研究方法相互补充，形成了一个较为系统的技术伦理框架。在针对ChatGPT的伦理研究中，Stahl等学者[22]将上述三种新兴技术伦理方法整合为一个综合框架，构成其核心方法论。这组研究框架也成为了本文的重要参考和初始框架。在此基础上，本文进一步分析生成式人工智能的技术机理，逐一识别生成流程中存在的潜在伦理风险，以初步建立伦理风险指标框架，旨在通过强调程序正义，确保算法系统符合伦理规范。

本文以ChatGPT为例，识别其生成流程中的潜在风险。OpenAI[23]揭示了ChatGPT生成流程的三个主要阶段：首先，监督微调（Supervised Fine-Tuning，SFT），利用人工标注问题和高质量答案微调模型；其次，奖励模型训练（Reward Model training），使用人工标注数据对生成的多个回答进行排序，形成训练数据，用于训练奖励模型；最后，采用近端策略优化（PPO）强化学习（Reinforcement Learning via Proximal Policy Optimization），通过奖励模型的打分更新预训练模型参数，形成策略梯度，从而提升模型的生成质量。由此可见，生成式算法通过训练模型来学习输入数据的分布并生成新数据，在此过程中，可能伴随多轮伦理风险。表1总结了算法流程与伦理风险之间的对应关系。

如表格所示，通过对生成式算法流程的梳理，我们可以识别如下潜在风险：

注入偏见的风险：在训练流程中，由于模型依赖于数据、训练模型和人类输入等多个组成部分，存在注入偏见的可能性。其中既包括人工智能训练师的偏见也包括来自数据集的偏见。OpenAI已经承认在ChatGPT的训练过程中存在一些偏见，比如训练师更倾向于生成详细答案[24]，尽管其已实施适度的政策规范ChatGPT的使用，但这些政策本身也可能存在偏见。因此，人工智能从业者需要开发评估方法，以衡量ChatGPT生

表1 生成式人工智能模型训练流程及其伦理风险

算法步骤	算法训练流程	算法流程中的伦理风险
步骤1：训练监督策略模型（搜集演示数据并训练监督策略）	1.从数据集中抽取一个提示	基于包含虚假信息的数据集可能导致生成虚假信息；基于互联网信息的数据集可能导致抄袭
	2.标记者演示所需的输出行为，生成回应	人类标注高质量输出导致偏见
	3.使用这些示范数据通过监督学习微调模型	监督学习微调导致偏见
步骤 2：训练奖励模型（收集比较数据并训练奖励模型）	4.抽取提示并生成多个不同的模型输出	模型显示从数据集中输出人们隐私信息的情况。
	5.标记者对输出进行打分排序	人类打分导致偏见
	6.用排序数据训练奖励模型	人类标注者对这些结果综合考虑给出排名顺序导致偏见
步骤3：采用强化学习来优化策略	7.采样一个新的问题	生成文本的性能受到使用的数据和训练模型的影响，可能导致抄袭、虚假信息、隐私、偏见、版权等一系列伦理问题。
	8.基于监督策略初始化 PPO 模型	
	9.该策略生成回答	
	10.奖励模型根据输出计算奖励	
	11.使用 PPO 和奖励更新策略	

成文本中偏见的程度。

虚假信息的风险：ChatGPT 是基于 GPT-3 开发的应用模型，后者作为拥有 1750 亿个参数的大型语言模型，基于从不同互联网来源（如网页、书籍、研究文章和社交对话）获取的海量数据进行训练[25]，使用数据和训练模型的选择直接影响了 ChatGPT 的性能。尤其在这些数据集可能包含大量虚假信息的情境下，ChatGPT 的操作可能会受到训练数据中虚假信息的引导。

操纵和误用的风险：生成式人工智能模型还可能被当作强大的操纵工具。[26]例如，当研究人员要求 ChatGPT 生成关于疫苗、新冠疫情、“国会山骚乱”、移民问题以及涉疆议题的虚假信息时，它都能够有效地完成这些需求。[27][28]因此，必须采取措施，防范这些工具被用于操纵和误导公众。这可能需要包括监管、技术改进和教育措施在内的多种策略，提高公众对生成式人工智能工具的认知，以更好地识别和应对潜在的虚假和错误信息。

此外，生成式算法流程带来的伦理风险还可能包括抄袭、版权、有毒内容输出等。我们识别出的潜在伦理风险构成了伦理指标体系的基本框架，这个框架的独特性在于它基于算法生成的机理，通过对生成式算法流程进行梳理和分析而得出，提示我们在算法的不同阶段介入伦理治理的切入点。例如，在预处理阶段，通过数据增强和公平数据生成来减少数据集中的偏见；在处理中阶段，通过调整模型参数来降低训练或推理过程中的偏见；在后处理阶段，通过审核和过滤生成内容来防止不公平内容的传播。[29]然而，生成式人工智能的伦理风险远不止于此，因此需要更系统严谨的方法，更全面地覆盖所有可能的伦理风险。本文将结合当前相关文献的伦理研究进行校验和补充，进一步构建完整的生成式人工智能伦理风险评估指标体系，以确保伦理指标的全面性和科学性。

（二）指标完善及体系建构

生成式人工智能伦理指标体系旨在扩大对生成式人工智能带来的伦理利益和挑战的认识，为了避免遗漏，本研究进一步检索了当前关于生成式人工智能伦理研究的相关文献。通过“生成式 AI（Generative AI）”、“大语言模型（LLM/Large Language Model）”、“伦理（Ethics）”、“风险（Risk）”等关键词检索文献，这一检索过程持续进行，直至未发现新的伦理问题为止。通过对从 Google Scholar、IEEE Xplore、Elsevier、Science-

表2 LDA主题挖掘结果示例

LDA Topics	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	ai	ai	user	ai	ai
2	privacy	generated	trustworthiness	ethics	factuality
3	security	content	systems	toxicity	bias
4	data	rights	builds	crucial	ethics
5	bias	property	confidence	responsible	toxicity
6	factuality	respected	ai	development	generated
7	major	intellectual	factuality	research	user

Direct、Arxiv等学术平台收集的100余篇相关论文进行筛选，同时剔除不相关的内容，这些最相关且重要的选定文献为本研究的伦理指标体系校验和补充提供了关键参考。

对这些语料文献进行的主题分析（Topic Analysis），有助于完善伦理指标以进一步构建完整的伦理指标体系。具体方法如下：首先运用LDA主题模型对研究文献进行了深入挖掘，识别出当前的主要伦理议题及关键词；在此基础上，通过关键词频率统计和词云图进行可视化分析，进一步明确了文献中高频出现的伦理风险领域；随后，本文对提取的关键词进行了聚类分析，归纳出二级指标，并最终构建多层级指标体系，用以全面评估生成式人工智能在多维度上的社会伦理影响。

1.LDA主题挖掘

首先运用潜在狄利克雷分配（Latent Dirichlet Allocation，LDA）模型对文献逐篇进行主题挖掘，LDA是一种自然语言处理中的主题模型技术，用于识别文档集合中的抽象主题。表2展示了通过LDA主题模型提取的五个主题，每个主题包含若干代表性关键词。例如，主题1主要围绕数据隐私、安全和偏见问题，关注事实准确性；主题2聚焦人工智能生成内容，涉及对知识产权和权利的讨论；主题3强调用户对系统的信任；主题4探讨了伦理、有害性以及人工智能系统开发与研究中的责任问题；主题5在人工智能与事实性、偏见和有害性之间的关系中更加突出用户体验的重要性。通过对这些主题的分析，我们能够识别多个维度上关于生成式人工智能伦理风险的关键议题。

2.关键词频率统计与词云图分析

在主题挖掘所提取的伦理关键词基础上，我们对文献逐篇进行关键词频率统计与词云图可视化分析，以进一步明确文献中频繁出现的伦理风险领域。如图1中的词云图通过直观的视觉效果，展示了文献中的伦理关注点，为伦理指标体系的构建与完善提供了有力支持。

3.指标完善

在前述分析基础上，我们对从语料文献中提取的关键词进行了系统分类，将其按照内在关联性聚类为具有代表性的伦理指标，以完善指标体系的构建。例如将诸如“Bias”（偏见）、“Discrimination”（歧视）、“Stereotypes”（刻板印象）等关键词归类到“偏见”这一伦理指标类别下，以便将所有相关的不同表述和现象集中在一起。通过对149个关键词的聚类分析，我们识别并归纳出了31个二级指标，这些指标涵盖了生成式人工智能研究中的多维度伦理、技术和安全性问题。该聚类过程不仅有助于厘清各个伦理议题的核心概念，还为研究者提供了一个结构化的统一分析框架，便于进一步深入探讨生成式人工智能技术中

图1 伦理关键词词云图示例

的伦理问题。

4.体系构建

最终，我们构建了一个由4个一级指标（技术、安全、社会责任、法律）、14个二级指标（虚假信息、可控性、透明性、技术风险、数据安全、系统安全、个人安全、社会公平、社会结构、价值对齐、环境影响、社会创造力、知识产权、合法性）及31个三级指标组成的多层级指标体系（表3），以全面评估生成式人工智能在多元维度上的社会伦理影响。值得注意的是，我们剖析了每一项伦理风险的算法根源及风险成因，这为后续针对性的风险治理奠定了基础。

（三）相关性检验

1.相关性检验

为验证我们构建的伦理指标体系的科学性和普适性，我们进一步将其与更广泛的研究文献进行了相关性检验。经过筛选，我们最终扩充至153篇高质量相关文献，提取其中的文本数据作为基础语料，进行关键词词频分析，以验证所构建的生成式人工智能伦理指标体系的普适性。结果显示，“偏见”（Bias）、“安全”（Security）、“信任”（Trust）、“公平性”（Fairness）和“隐私”（Privacy）等关键词词频最高，而虚假信息类别以“深度伪造”（Deepfake）、“虚假”（Fake）等关键词分散显示，尽管个体词频并未达到最高水平，但它们总体揭示了虚假信息传播风险在相关研究中占据重要地位，上述高频词汇反映了当前研究中最为集中的伦理关注点；“攻击”（Attack）、“监管”（Regulation）和“透明度”（Transparency）等关键词位居第二梯队，表明应对安全威胁以及建立健全的监管框架是研究中的重要挑战。此外，“价值对齐”（Alignment）、“知识产权”（Copyright）和“环境”（Environment）等主题词频也在相关研究中具有一定的关注度。通过关键词频检验，我们发现所构建的伦理指标体系能够全面覆盖这些关键词，初步验证了该体系的全面性和适用性。

2.指标赋权

指标赋权能够科学地反映不同指标在体系中的相对重要性，反映学术界和行业内部对不同伦理问题的关注度。我们采用归一化方法对伦理指标进行了权重分配。归一法是一种常用的权重赋值方法，通过标准化数据，将所有指标的权重归一化（使权重的总和为1），使得不同维度或不同单位的数据可以在同一尺度上进行比较。这种方法在多指标综合评价中尤为常见。我们首先以三级指标的词频为基础计算其初始权重，并对其进行归一化处理。运用归一化法中的（0，1）法，将各条目词频数转化为1以下的小数，计算出各条目的权重值；接着，将属于相同二级指标下的三级指标权重相加，以确定二级指标的权重。通过为伦理指标赋予相应的权重，我们最终构建的“生成式人工智能伦理指标体系”如表4所示。

五、主要发现：生成式人工智能的伦理光谱

在前文构建的生成式人工智能伦理指标体系中，不同伦理议题在权重上的差异反映了当前伦理研究中的核心关切，折射出当前生成式人工智能伦理研究的整体光谱。解读指标体系中的不同权重层级的伦理风险背后深层次的技术逻辑及风险成因，有利于在后续的治理过程中提出更加精准的应对策略。

（一）高权重的核心伦理议题

表3 生成式人工智能伦理风险指标体系

一级指标	二级指标	三级指标	伦理风险的算法根源
技术	虚假信息	虚假信息	生成式人工智能的可靠性受训练数据、模型架构影响，表现出一定的概率性、随机性，包含虚假信息的数据集可能导致生成虚假信息。
	可控性	可控性	算法黑箱性及系统易攻击性，使得开发者难以预判和控制模型的具体生成行为。
		可追溯性	生成内容过程中缺乏完整的溯源机制，难以追溯生成过程。需要对数据生成过程及后续传播追踪，以便监控涉及安全的行为[30]
	透明性	透明性	黑箱模型的复杂性使其内部操作难以被外部理解和解释。
		解释性	高度复杂的模型结构和参数使得模型输出难以向他人解释。
	技术风险	技术滥用	算法可能被滥用于深度伪造等，超出预期应用范围。
		技术垄断	生成模型的高开发成本和资源需求导致少数大公司垄断。
		技术依赖	人类可能过度依赖技术解决方案，进而存在技能退化的风险
安全	数据安全	数据安全	算法训练过程中使用的数据集可能包含敏感信息，若不加以保护，可能导致数据泄露。
		信息安全	算法安全性漏洞可能被恶意利用，导致信息泄露。
	系统安全	侵入性	生成式人工智能系统存在“越狱”风险，黑客使用提示注入等技术规避安全防护机制[31]，或导致输出内容被篡改或恶意利用。
		操纵性	算法可能被操纵生成特定类型信息，以影响公众意见或行为。
		鲁棒性	对输入数据微小变化表现出的不稳定性导致算法输出不可预测。
	个人安全	隐私风险	大量真实数据被用于算法训练，当数据集包含私人数据时，模型可能复制敏感信息，并从生成内容中泄露隐私。
		有毒性	算法“垃圾进，垃圾出”的原理，导致大语言模型生成不道德、欺诈、色情或其他有毒性内容。
		伤害性	算法生成的内容可能包含暴力等伤害性内容，影响身心健康。
社会责任	社会公平	偏见	训练数据可能携带各种社会偏见，如种族主义、性别歧视、残疾人歧视等，导致生成内容带有偏见。
		公平性	算法的决策过程可能不公平，导致某些群体受到的歧视或忽视。
		正义	算法生成的内容影响公众对正义的理解和判断。
		包容性	算法可能忽视社会多样性，生成内容缺乏包容性。
	社会结构	权力垄断	算法的使用、控制集中在少数人手中，可能导致权力分配不均衡。
		劳动力置换	算法自动化的广泛应用可能取代人类劳动，导致失业问题。
		数字鸿沟	算法技术的复杂性和高成本可能扩大技术接入的不平等，比如发展中国家或经济困难群体无法平等获取生成式人工智能。[32]
	价值对齐	价值观对齐	系统通过人类反馈学习人类的价值观[33]，但选择哪些价值观作为系统的指导标准存在争议，可能导致价值观对齐问题。
	环境影响	环境影响	算法训练所需的计算资源消耗大量能源，对环境造成负面影响。
		经济影响	算法带来的经济效益可能分配不均衡，影响社会的经济结构。
	社会创造力	创造力	过度依赖算法生成内容可能抑制人类的创造力，削弱原创性。
法律	知识产权	知识产权	算法生成的内容通过抓取互联网数据进行训练，可能在无意中涉及现有作品的版权问题，导致知识产权纠纷。
		剽窃	算法生成的内容可能未经授权使用他人作品，构成剽窃行为。
	合法性	网络犯罪	算法生成的内容可能被恶意用于诈骗等网络犯罪活动。
		合规性	算法的开发应用可能面临复杂的法律法规要求，难以完全合规。

在伦理指标体系的三级伦理指标中，权重超过或接近0.10的高权重指标包括“虚假信息”“信息安全”“公平性”“偏见”和“隐私风险”。这些高权重的伦理指标反映了当前学术界对这些问题的高度关注，构成了生成式人工智能伦理研究中的核心议题。

1.虚假信息

生成式人工智能模型的性能和生成文本的可靠性深受其训练数据和模型架构的影响，表现出一定的概率性和随机性，这种机制导致了生成、传播虚假信息的潜在风险。美国新闻可信度评估和研究机构NewsGuard指出，生成式人工智能工具可能会以前所未有的规模传播虚假信息。经测试表明，ChatGPT从NewsGuard的专有错误信息数据库中获得了100个样本，为100个已确定为错误论点中的80个生成了错误叙述[34]，凸显了其在生成虚假信息方面的风险。有研究人员从SEO排名较高的软件中随机选择了Simplified App、AI Writer、Copymatic、Sassbook Writer和Write Sonic五款软件，用于创建关于“人工智能的增长”主题的内容[35]，以评估这些软件生成的内容的质量和可靠性。结果显示，这5款软件的准确性和可靠性最高为37.5%（Sassbook Writer），而最低仅为12.5%（Copymatic），这项实证研究表明了生成式人工智能在准确性和可靠性等方面面临挑战。此外，深度伪造、假新闻等衍生风险可能进一步导致信息扭曲、误导公众，甚至引发社会混乱。

2.信息安全

根据前文对生成式人工智能模型训练流程的分析可知，生成式人工智能模型依赖人类提供的提示词（Prompt）来理解并生成相关内容，这为黑客通过提示注入等技术手段绕过系统安全防护提供了机会，从而得以诱导模型生成含有偏见或误导性的结果，暴露了生成式人工智能在安全性上的潜在漏洞。生成式人工智能系统的“越狱风险”日渐成为讨论焦点，而后门植入和模型投毒技术[36]等攻击手段，进一步加剧了系统的脆弱性。这种脆弱性不仅可能导致信息泄露，还为恶意操纵提供了途径。更为严重的是，即便具备有限技术能力的网络犯罪分子，亦可利用这些人工智能系统获取黑客工具的相关信息。这类系统也可能被用于实施“软战”策略，以实现特定的政治或商业目的，凸显了加强系统安全性和防护措施的紧迫性。

3.公平性和偏见

生成式人工智能的偏见问题主要源于其训练数据和算法设计本身，训练数据中所携带的偏见可能在模型生成的内容中被延续并放大。大量案例表明，种族主义、性别歧视以及对残疾人的歧视等常通过数据进入模型并在生成内容中显现。在招聘、贷款审批、刑事司法等对公平性要求极高的应用领域[37]，这类偏见所带来的负面影响尤其深远。还有研究指出，大型语言模型可能将数据集中的社会不平等关系嵌入自身，进而反映在其输出内容中，无意中强化并加剧现有的权力结构。例如，ChatGPT的输出往往倾向于以中产阶级白人男性的视角呈现，在多样性和包容性上无法充分代表社会整体。[38]而当新一代生成式模型以先前模型生成的合成数据作为训练基础时，偏见可能会被进一步放大和固化。[39]因此，如何避免生成式人工智能在算法设计和模型应用中产生存在偏见或不公平的结果，已成为伦理讨论的核心关注点。

4.隐私风险

早期模型已出现从数据集中泄露个人信息的案例，随着生成式人工智能模型的规模不断提升，大量真实数据被用于模型训练与内容生成，这使得模型可能“记忆”某些私人数据，带来了潜在的隐私风险。在特定条件下，这些敏感信息可

表4 归一化赋权的生成式人工智能伦理风险指标体系

一级指标	二级指标	二级指标权重	三级指标	三级指标权重	一级指标	二级指标	二级指标权重	三级指标	三级指标权重
技术	虚假信息	0.155	虚假信息	0.155	社会责任	社会公平	0.234	偏见	0.104
	可控性	0.044	可控性	0.043				公平性	0.109
			可追溯性	0.001				正义	0.006
	透明性	0.079	透明性	0.061				包容性	0.014
			解释性	0.018		社会结构	0.013	权力垄断	0.002
	技术风险	0.01	技术滥用	0.008				劳动力置换	0.004
			技术垄断	0.001				数字鸿沟	0.007
			技术依赖	0.001		价值对齐	0.013	价值观对齐	0.013
安全	数据安全	0.167	数据安全	0.044		环境影响	0.036	环境影响	0.032
			信息安全	0.123				经济影响	0.004
	系统安全	0.08	侵入性	0.065		社会创造力	0.003	创造力	0.003
			操纵性	0.005	法律	知识产权	0.026	知识产权	0.024
			鲁棒性	0.010				剽窃	0.001
	个人安全	0.109	隐私风险	0.099		合法性	0.031	网络犯罪	0.001
			有毒性	0.007				合规性	0.030
			伤害性	0.003					

能被无意复制或公开，进而导致隐私泄露。研究表明，训练数据的规模扩大，隐私风险也随之增加。此外，用户在与生成式人工智能模型互动过程中，难免分享某些个人信息，这可能引发数据泄露和未经授权访问个人识别信息的风险，进而导致社会结构性的“隐私剥夺感”（Privacy deprivation）。[10]因此，确保信息的保密性和安全性，并清晰展示信息被使用的方式尤为重要。

（二）中等权重的次核心伦理议题

中等权重的伦理指标（权重在0.01到0.09之间）包括“价值观对齐”“知识产权”“环境影响”“合规性”“侵入性”“透明性”“可控性”“解释性”“鲁棒性”和“包容性”等，这些议题虽然紧迫性略低于前者，但在伦理治理中仍需重点考虑。

1.价值观对齐

在生成式人工智能系统的训练过程中，模型通过人类反馈、观察和讨论等方式学习和适应人类的价值观，而这一技术机理也带来了价值观对齐方面的挑战。生成式人工智能的价值对齐原则总体上旨在训练系统，确保其行为与人类价值观保持一致[11]，但在跨文化语境中，如何定义和选择适合作为生成式人工智能指导标准的价值观仍然存在较大争议。生成式人工智能的训练过程中可能反映出不同社会的价值观冲突（比如通过人类打分的反馈机制等），导致生成的内容无法完全对齐某一特定文化或社会体系的价值观并与其方向偏离。

2.知识产权

生成式人工智能模型通过从互联网抓取大量数据进行训练，其训练过程依赖于已有的内容资源，这种技术逻辑使得模型在生成新内容时，容易无意间复制受版权保护的材料，导致版权侵权行为。自动生成功能也引发了关于学术成果问责和作者身份的争议，可能导致出现学术不端行为，并削弱科研成功的可信度。[12]生成式人工智能是否能够被视为合著者，以及它如何影响对学生的评估等，都是需要严肃对待的议题。对于人工智能辅助或参与撰写的手稿，需要厘清文本的所

有权应归属于训练数据的原作者，还是生成式人工智能的所属公司，或是使用该系统进行创作的科研人员。这些问题需要通过深入的法律讨论和设立明确的法规加以解决。[43]

3.环境影响

依赖于高性能计算机集群的生成式算法通常需要大量计算资源，尤其是大规模预训练模型（如GPT系列、DALL-E等）的训练和运行，这使得生成式人工智能技术的发展伴随着巨大的环境成本。训练一个大型生成模型可能消耗数百兆瓦时的电力，这不仅带来了直接的能源消耗问题，还间接导致了大量的碳排放。有研究表明，训练一个大型模型产生的二氧化碳排放量约相当于一辆汽车生命周期排放量的1到10倍。[44]人工智能模型越复杂，所耗计算资源越多，其对能源消耗和生态系统的潜在影响可能会愈发严重，因而生成式人工智能技术的碳排放问题应该成为政策和监管关注的重点。

4.合规性

合规性（Compliance）是指在生成式人工智能系统中，确保遵守相关法律、法规及道德标准的要求。“垃圾进，垃圾出”的原理反映了生成式人工智能系统输出与其训练数据质量之间的密切关系。生成式大语言模型的技术逻辑决定了其容易被滥用于生成不道德、欺诈、色情等违法有害内容，乃至被犯罪分子用以进行创建虚假身份或编写网络钓鱼信息等社会工程攻击（Social Engineering Attacks）[45]，进而引发严重的社会后果。建立清晰的法律框架，确保生成式内容的合法性和合规性，已成为生成式人工智能发展的重要议题。

（三）低权重的潜在伦理议题

当前研究中权重相对较低的伦理议题（权重低于0.01）包括：创造力、权力垄断、技术依赖、技术滥用、有毒性、数字鸿沟、正义、操纵性、劳动力置换、经济影响、伤害性、可追溯性、技术垄断、剽窃、网络犯罪等。例如创造力问题是由于生成式人工智能系统仅限于结合现有信息，限制了在推动创新方面的潜力；技术依赖是指生成式人工智能可能导致组织过度依赖技术解决方案，忽视了人类在学习和决策中的不可替代性，进而带来技能退化的风险；[46]与此同时，由于开发生成式大模型成本高昂，少数大型人工智能实验室对基础模型的开发形成了技术垄断现象。此外，生成式人工智能的快速发展将会对劳动力市场造成结构性影响，尽管其推动了提示工程师等新兴职业的出现，但大量传统岗位面临被替代的风险，进一步加剧了社会经济不平等，这也是劳动力置换与经济影响的核心关切。尽管目前对以上伦理议题的关注度相对较低，但仍需防微杜渐，确保技术发展过程中的潜在风险能够得到及时有效的治理。

综上，本文通过构建伦理指标体系，为生成式人工智能伦理研究提供了统一的框架，还通过权重分布明确了各伦理议题的重要性和优先级。其中的高权重议题凸显了亟待解决且得到重点关注的关键问题，而中低权重的伦理议题，如价值观对齐、环境影响、技术依赖和权力垄断等，则为长远的技术发展发出了预警信号。通过深入剖析不同权重议题背后的技术逻辑和风险成因，我们能够为伦理治理提供更具针对性和前瞻性的政策建议，推动生成式人工智能可持续发展。

六、生成式人工智能伦理风险的针对性治理

前述伦理指标体系勾勒了生成式人工智能所面临的伦理风险全景，并探讨了不同风险等级背后的技术逻辑及其成因。基于上述分析，本节将提出针对性的治理策略，重点讨论技术治理与社会治理相结合的综合方案及其具体实施路径。

（一）技术治理

技术治理是规避伦理风险的第一道防线，旨在基于生成式人工智能的技术逻辑，通过优化技

术本身，减少系统运行中的伦理风险。面对前述伦理指标体系中的隐私风险、虚假信息、知识产权、环境影响等重要伦理议题，建议针对其风险成因，采取以下关键技术治理措施：

1.隐私保护技术

分析风险成因可知，隐私风险主要源于生成式人工智能对真实数据的依赖性，尤其是在模型训练过程中可能“记忆”并复制私人数据，导致隐私泄露风险增加。对此，有研究人员提出了机器遗忘（Machine Unlearning）[47]的解决方案。通过让模型“遗忘”特定的训练数据，防止敏感信息的泄露，并减少重新训练模型的必要性。类似的，有学者提出了记忆拒绝（Memorization Rejection），放弃与训练数据近似重复的生成数据。而 Stable Diffusion 的轻量级模型补丁 Forget-Me-Not[48]，可以有效移除包含特定身份的概念，避免生成任何具有该身份的面部照片。此外有研究者提出可以通过生成式人工智能技术替换敏感信息来保护用户隐私。例如，研究人员只需使用一个人名提示词就可以轻易生成该人的面部图像，该图像几乎与真实训练样本相同。而如果通过用虚拟内容替换敏感内容来保护隐私，则生成的图像与真实图像具有不同的身份，可以阻止未经授权的识别。[49]

2.虚假信息检测

在技术层面，生成式人工智能模型依赖于概率分布进行内容生成，导致生成内容可能包含虚假或误导性信息。应对这一风险的关键方案是确保用户能够辨别数据是否由人工智能生成。生成检测（Generative Detection）技术能够有效区分真实数据与人工智能生成的数据，帮助防止虚假信息的传播。例如，华为团队发布的GenImage项目[50]被称为“AIGC时代的ImageNet数据集”，该项目涵盖数百万张图像，并采用了最先进的生成器，有助于快速进行区分。此外，生成归因（Generative Attribution）技术[51]通过追踪生成数据的来源，进一步增强了对虚假信息的防范能力。这些技术手段，为监控生成式人工智能的虚假信息传播提供了强有力的工具，降低了信息误导的社会风险。

3.版权保护技术

针对人工智能生成内容的抄袭与版权保护问题，技术上可以利用语言模型的能力构建文本检测器（Text Detection）来识别人工智能生成的文本，进而降低抄袭风险。普林斯顿大学开发的文本检测工具GPTZero即是此类技术的典型案例。此外，数字水印（Digital Watermarking）和区块链（Blockchain）技术也提供了可靠的版权保护机制。在生成模型或生成数据中嵌入数字水印，可以追踪内容来源，实现版权保护。通过区块链记录版权信息，能够确保生成数据与特定版权持有者的关联性，减少版权纠纷的发生。[52]同时，生成模型的开发者和用户也需遵守相关版权法律，避免使用受保护的内容进行训练或生成类似作品，以降低版权侵权的风险。

4.可追溯性机制

可追溯性机制旨在确保生成内容出现问题时可以准确定位生成模型的开发者或用户，并根据相关法律对其进行追责。生成内容的责任追究可以通过生成归因（Generative Attribution）技术实现，即确定生成数据的来源和模型。区块链技术同样可以实现生成数据的可追溯性——每个生成数据被记录在区块链中的一个区块，并与相应的交易或生成过程相关联。用户和监管者能够了解生成数据的来源和完整的生成路径。[53]

5.环境影响最小化

随着人工智能模型的复杂性增加，其所需的计算资源显著提升。技术治理应当涵盖开发更高效的计算方法，以减少碳排放和能源消耗。有研究者提出“绿色编码实践”（Green Coding Practices）[54]的概念，包括算法优化及代码简化等，旨

在通过在软件开发过程中采用减少能源消耗和资源浪费的编码策略来降低环境影响。与此同时，可进一步探索可再生能源在数据中心的应用，以有效降低环境负担。

（二）社会治理

在技术治理之外，本文提出“三套机制—三方监督”的社会治理框架，以全面保障生成式人工智能技术的负责任使用。

1.坚持人工核查机制

尽管生成式大模型能够有效加速总结、评估和审查过程，人工核查仍是防止信息失实和道德偏差的关键。OpenAI在发布ChatGPT时，曾采取多项限制措施，如阻止其访问互联网并安装过滤器，以避免生成敏感或有害内容，而这些技术手段仍需要依赖人工审核员来标记并过滤有害信息[59]，确保内容的可靠性和准确性。

2.制定问责机制

建立明确的问责机制要求研究人员应当披露生成式人工智能在研究中的具体作用，并对生成的内容、结果、数据、代码和参考文献负责，确保技术应用在学术研究中保持透明和负责任，避免因技术使用不当而损害研究的可信度。

3.建构跨学科伦理审查机制

生成式人工智能的应用必须经过跨学科的伦理审查，技术专家、伦理学家、法律专业人士应共同参与。这种多元化的审查机制可以提供全面的风险评估，确保技术应用的合规性和合法性。

4.建立第三方监督机构

独立的第三方监管机构可以对生成式人工智能的开发和应用进行全方位监督。同时，公众的参与能够确保技术决策符合社会利益，防止技术滥用，促进公平和正义。

综上，生成式人工智能的伦理风险治理需要综合技术治理与社会治理，形成一个多层次、多角度的伦理风险应对体系。在技术层面，通过改进算法、优化数据以及提升系统透明度，减少生成式人工智能潜在的伦理风险；在社会层面，通过完善法律政策、建立问责和监督机制，确保技术应用符合社会伦理要求。最终，技术与社会的双重治理相辅相成，为生成式人工智能技术的可持续发展提供保障。

七、结语

生成式人工智能的快速发展伴随着诸多伦理风险。本文围绕这些风险展开分析，通过创新性的方法论，科学系统地构建伦理风险指标体系，并提出治理方案。首先，通过深入剖析生成式人工智能的算法机理，识别潜在的伦理风险；其次，构建了涵盖多层级、多维度的伦理指标体系，该体系不仅有助于确保生成过程符合伦理规范，还通过减少主观因素的影响，增强了其科学性与严谨性。最后，本文通过剖析伦理风险背后的技术逻辑，提出了针对性的治理路径。

本文的特别贡献还在于从算法机理的技术层面揭示了伦理风险的根源。通过这种将技术逻辑与伦理框架结合的综合方法，本文为生成式人工智能的伦理风险治理提供了更加精确和可操作的参考路径，弥补了现有研究中对技术机制与伦理规范互动分析的不足，进一步拓展了该领域研究的深度和广度。

[本文系国家社科基金重大项目“5G时代新闻传播的格局变迁与研究范式转型”（项目编号：21&ZD325）、上海市哲学社会科学规划课题项目“融媒体时代基于AI社交知识图谱的舆论智慧引导机制研究”（项目编号：2021BXW009）、新华社媒体融合生产技术与系统国家重点实验室项目“智能媒体视角下伦理算法与社会责任评价指标体系应用研究”（项目编号：SKLMCPTS202103010）的研究成果]

[陶婷婷：上海交通大学媒体与传播学院助

理研究员；李本乾（通讯作者）：上海交通大学智能传播研究院院长、上海交通大学媒体与传播学院特聘教授、博士生导师]

注释：

①②㉑㉕㉗㊳㊷㊻ Dwivedi, Yogesh K., et al., "Opinion Paper: 'So What If ChatGPT Wrote It?' Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy", International Journal of Information Management, vol. 71, 2023, 102642.

③European Commission., Proposal for a Regulation on a European Approach for Artificial Intelligence. European Commission, 2021-04, https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence, 2024-06-24.

④M ü ller, Vincent C., "Ethics of Artificial Intelligence and Robotics." in Edward N. Zalta, eds., The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, 2020-04, https://plato.stanford.edu/archives/fall2020/entries/ethics-ai/, 2024-06-24.

⑤Vesnic-Alujevic, Lucia, et al., "Societal and Ethical Impacts of Artificial Intelligence: Critical Notes on European Policy Frameworks", Telecommunications Policy, vol. 44, no. 6, 2020, 101961.

⑥Shmueli, Galit, et al., "How Can IJDS Authors, Reviewers, and Editors Use (and Misuse) Generative AI?", INFORMS Journal on Data Science, vol. 2, no. 1, 2023, pp.1-9.

⑦Susarla, Anjana, et al., "The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems", Information Systems Research, vol. 34, no. 2, 2023, pp.399-408.

⑧⑩⑬⑱㉒Stahl, Bernd C., and David Eke., "The Ethics of ChatGPT -Exploring the Ethical Issues of an Emerging Technology", International Journal of Information Management, vol. 74, no.4, 2024, 102700.

⑨ Weidinger, Laura, et al., "Ethical and Social Risks of Harm from Language Models", arXiv, 2021-12, http://arxiv.org/abs/2112.04359, 2024-06-24.

⑪⑫⑮Mougan, C., and J. Brand., "Kantian Deontology Meets AI Alignment: Towards Morally Grounded Fairness Metrics", arXiv, 2024-02, http://arxiv.org/abs/2311.05227v2, 2024-06-28.

⑭Gray, Paul, and Anat Hovav., "The IS Organization of the Future: Four Scenarios for 2020", Information Systems Management, vol. 24, no. 2, 2007, pp.113-120.

⑯Gaver, William W., "Technology affordances",Proceedings of the SIGCHI conference on Human factors in computing systems, 1991, pp.79-84.

⑰李本乾、姬雁楠：《身临其境与社交互动：可供性视角下VR使用意愿的影响模式研究》，《现代传播(中国传媒大学学报)》2024年第1期，第129-137页。

⑲Brey, Philip A. E. "Anticipating Ethical Issues in Emerging IT", Ethics and Information Technology, vol. 14, no. 4, 2012, pp. 305-317.

⑳Wright, David. "A Framework for the Ethical Impact Assessment of Information Technology", Ethics and Information Technology, vol. 13, no. 3, 2011, pp. 199-226.

㉑Stahl, Bernd C., et al. "Ethics of Emerging Information and Communication Technologies: On the Implementation of Responsible Research and Innovation", Science and Public Policy, vol. 44, no. 3, 2017, pp. 369-381.

㉓㉔ OpenAI, "ChatGPT: Optimizing Language Models for Dialogue", 2022-12, https://openai.com/blog/chatgpt/, 2024-06-28.

㉖Kshetri, Nir., "ChatGPT in Developing Economies", IT Professional, vol. 25, no. 2, 2023, pp. 16-19.

㉘Klepper, David., "It Turns Out That ChatGPT Is Really Good at Creating Online Propaganda: 'I Think What's Clear Is That in the Wrong Hands There's Going to Be a Lot of Trouble'", Fortune, 2023-01, https://fortune.com/2023/01/24/chatgpt-open-ai-onlinepropaganda/, 2024-06-29.

㉙ Li, Sheng, et al., "Trustworthy AI-Generative Content in Intelligent 6G Network: Adversarial, Privacy, and Fairness", arXiv, 2024-05, http://arxiv.org/abs/2405.05930, 2024-06-29.

㉚㊼㊾㊶㊷㊸Wang, Tao, et al., "Security and privacy on generative data in aigc: A survey",arXiv, 2023-12, https://arxiv.org/abs/2309.09435, 2024-06-30.

㉛ ㉜ ㉝ Hagendorff, Thilo., "Mapping the Ethics of Generative AI: A Comprehensive Scoping Review", arXiv, 2024-02, https://arxiv.org/abs/2402.08323v1, 2024-06-29.

㉞㊲ Cao, Yuan, et al., "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT", arXiv, 2023-03, http://arxiv.org/abs/2303.04226, 2024-06-29.

㉟Ravichandran, K., and Senthil Kumar Ilango., "A Preliminary Analysis of the Quality of the Content Produced by AI Bots Using AI in Content Generation",International Journal of Intelligent Systems and Applications in Engineering,vol.11, no.5s,2023, pp.585-590.

㊱㊴㊶㊺ Hagendorff, Thilo., "Mapping the Ethics of Generative AI: A Comprehensive Scoping Review", arXiv, 2024-02, https://arxiv.org/abs/2402.08323v1, 2024-06-29.

㊵杨喜喜、李本乾：《结构性剥夺：算法受众隐私剥夺感的生成逻辑》，《山东师范大学学报(社会科学版)》2023年第6期，第125-138页。

㊸Van Dis, Emiel A. M., et al., "ChatGPT: Five Priorities for Research", Nature, vol. 614, no. 7947, 2023, pp. 224-226.

㊹㊹Vartziotis, Thomas, et al., "Learn to Code Sustainably: An Empirical Study on LLM-Based Green Code Generation", arXiv, 2024-03, http://arxiv.org/abs/2403.03344, 2024-06-30.

㊽E. Zhang, K. Wang, X. Xu, Z. Wang, and H. Shi, "Forget-me-not: Learning to forget in text-to-image diffusion models", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp.1755-1764.

㊿ Zhu, Mingjian, et al., "Genimage: A million-scale benchmark for detecting ai-generated image", in Advances in Neural Information Processing Systems, vol.36, 2024.

55 Stokel-Walker, Chris, and Richard Van Noorden., "What ChatGPT and Generative AI Mean for Science", Nature, vol. 614, no. 7947, 2023, pp. 214-216.

（责任编辑：王雨阳）