生成式人工智能的算法伦理难点分析与探索

施敏，杨海军
上海市互联网信息办公室，上海 200032

摘要
自2022年下半年以来，生成式人工智能技术和产业快速发展。聚焦生成式人工智能技术所用的生成式算法，结合相关法规规范，提出“生成式算法三定律”伦理原则。结合其技术特点，对实践中存在的伦理难点开展分析，并初步探索并提出解决框架。

关键词
生成式人工智能；大语言模型；生成式算法；算法伦理

中图分类号：TP391 文献标志码：A doi：10.11959/j.issn.2096-0271.2025013

Analysis and exploration of algorithmic ethical difficulties in generative artificial intelligence

SHI Min, YANG Haijun
CyberSpace Administration of Shanghai, Shanghai 200032, China

Abstract
Since 2022, generative artificial intelligence technology and industry have been developing rapidly. This paper focused on the generative algorithms used in generative AI technology, and proposed the "Three Laws of Generative Algorithms" ethical principles in combination with relevant regulations. Combined with its technical characteristics, it analyzed the ethical difficulties that exist in practice, and then explored some preliminary solution frameworks.

Key words
generative artificial intelligence, large language model, generative algorithm, algorithmic ethics

0 引言

自2022年下半年以来，生成式人工智能技术和产业快速发展。根据《互联网信息服务算法推荐管理规定》《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》中的有关定义，生成式人工智能技术（具有文本、图片、音频、视频等内容生成能力的模型及相关技术）用的是生成类算法。本文聚焦生成式算法，结合相关法律法规、伦理规范，提出“生成式算法三定律”的伦理原则；同时，结合其技术特点，分析在实践中存在的伦理难点，并初步探索解决框架。

1 相关法律法规对生成式算法的伦理要求

根据《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》3部上位法、上述3部算法相关法规、我国提出的《全球人工智能治理倡议》和《科技伦理审查办法》等，针对面向公众提供服务的生成式算法的合规和伦理要求，形成“生成式算法三定律”（12条指引）。

1.1 第一定律：生成式算法应“以人为本”[1]，保障人的隐私和合法权益

（1）算法应保障所有用户的知情权、选择停止权和涉个人信息的删除权。

（2）算法在训练、优化、提供服务中使用、生成的数据，涉个人信息的，应取得个人同意或符合法律法规规定，尊重他人肖像权、名誉权、荣誉权、隐私权和个人信息权益，不得危害他人身心健康。

（3）在做到第一点的基础上，算法应采取措施充分尊重并保护特殊群体的合法权益，如防范未成年人过度依赖或沉迷、对老年人的智能化适老服务和防范电信网络诈骗等。

（4）算法应造福于人类，增进人类共同福祉，保障社会安全，尊重人类权益。

1.2 第二定律：生成式算法应遵循向上向善、公平公正原则

（1）算法生成的内容符合和平、发展、公平、正义、民主、自由的全人类共同价值和所在国家、地区的价值观要求。不得利用算法生成各类法律、法规禁止和违背伦理道德的内容。不得利用算法操纵舆论、传播虚假信息。鼓励生成积极健康、向上向善的优质内容。对图片、视频等生成内容应予以标识。

（2）在算法的设计、优化与应用中，应采取有效措施防止民族、信仰、国别、地域、性别、年龄、职业、健康等各类歧视。

（3）不得利用算法实施侵犯知识产权、违背商业道德、垄断和不正当竞争等行为。

（4）明确和公开算法服务的适用人群、场合、用途，指导使用者科学理性认识和依法适用。制定并公开算法的相关基本原理、目的意图和运行机制。

1.3 第三定律：生成式算法应不断提升安全性、可靠性、可解释性和自主性

（1）应采取有效措施，保障与算法相关的模型、数据、基础设施、应用等安全，提供安全、稳定、持续的服务。防范对生成式人工智能技术的恶用、滥用。

（2）基于服务的类型特点，应采取有

效措施，不断提升算法的可解释性和可预测性，提升服务透明度，提高生成内容的准确性和可靠性，确保生成式人工智能始终处于人类控制之下，打造可审核、可监督、可追溯、可信赖的技术。

（3）研发、提供具有舆论属性或社会动员能力的算法模型，应建立健全算法机制机理审核验证、科技伦理审查、安全评估、应急处置、投诉举报等管理制度和技术措施。

（4）鼓励生成式算法及相关基础技术的自主创新。应使用具有合法来源的基础模型，采用安全可信的软件、工具和数据资源等。

2 生成式算法伦理实践中存在的难点分析

生成式模型有三大要素：高并发大算力、海量语料数据和复杂集成的算法。其中，Transformer架构生成式算法的主要技术包括词向量的表示、编码器-解码器架构、自注意力机制、预训练和微调、多任务学习、分布式语义等[2]。基于上述技术特点，逐一分析实践中可能存在的伦理问题和难点。

2.1 机器幻觉造成信息误导、歧视偏见，价值观参差不齐，违背向上向善、公平公正原则

大语言模型的机器幻觉通常是指模型在生成文本时，产生了不符合事实、逻辑或常识的内容。出现机器幻觉的技术原理，主要有4个方面。一是深度学习模型的局限性。模型会学习到数据中的偏见和错误信息，并在生成文本时反映出来。二是过度泛化。处理训练数据中，模型并非真正理解文本含义，而是试图找到一种简单方法来生成文本，这些文本在训练数据中出现频率较高、但并不符合事实或逻辑。三是优化目标不一致。训练目标是最大化生成文本的概率，生成文本时，更注重提高文本的概率而非确保其准确性和一致性。四是训练数据不足。模型无法学习到足够的信息和知识[3]。数据如存在偏见或歧视，生成内容也可能继承，如训练数据中男女职业分布不均衡，生成内容可能也会体现出这种不平衡。此外，超参数设置、自注意力机制过于复杂也可能导致机器幻觉。

机器幻觉问题可能带来的伦理问题有：误导公众，即生成不符合事实、逻辑或常识的内容，可能误导用户特别是青少年传播错误信息，甚至影响社会稳定，违背向上向善原则；公平性缺失，即生成带有偏见歧视的内容，违背公平公正原则，违背生成式算法“第二定律”；仅完成预训练、SFT的模型，生成内容与人类共同价值观、所在国家和地区价值观可能差异较大。此外，考虑到大语言模型的复杂性、海量文本“千人千面”的生成机制，在人机互动场景中，采用常规“机审+人审”方法对生成内容实现有效过滤监管也存在难度。

2.2 持续追求模型的能力提升与资源耗费、模型可解释性下降等问题之间的矛盾

自GPT3开始，千亿模型时代来临，国内各企业大模型也基本是千亿级别。参数数量通常与模型的大小、复杂性和表达能力有关。更多参数意味着模型可学习到更复杂的特征和模式。虽然更大的模型通常具有更强的表达能力和泛化能力，能处理更复杂的任务，但参数是否越多越好？过多的参数至少存在3个风险。一是过拟

合风险。模型可能会记住训练数据中的噪声和偏差，导致对未知数据的性能下降。二是模型复杂度太高带来的风险。模型会难以解释和调试，遇到问题难以定位和解决。三是训练和推理的计算资源和时间成本过高。按相关文献，训练所需算力可粗略估算为：参数量×批大小/学习率[4]。一个1 000亿参数模型训练所需算力，假设批大小为64，学习率为0.001，训练所需算力约为16 000 TFLOPS，换算成A100算力约821张卡，事实上考虑硬件冗余、通信开销，可能要超过千张卡并行算力，即业界所说千亿参数模型需要千卡算力。计算资源和时间的增加会造成训练和推理过程中需要更多能源，这可能导致碳排放量增加，对环境造成负面影响。

不断追求高能力、大参数可能造成算法模型可靠性、可解释性下降，违背算法“第三定律”。碳排放量增加会加剧全球气候变暖，违背造福人类“以人为本”的“第一定律”。

2.3 个人信息、重要数据泄露风险点增多，违背保障人的隐私和合法权益原则

基于笔者在《大语言模型的数据隐私保护难点分析与探索》中的观点，预训练收集的海量语料数据中含有大量个人信息和重要数据，深度学习技术提升属性预测能力使模型成为“社工利器”，各类组织和人群滥用、恶用算法模型实施违法犯罪，此外算法实现、优化、服务中的各类安全风险均可能导致个人信息、商业机密甚至国家安全数据的泄露风险增大[5]，侵犯他人隐私权、名誉权等合法权益。数据来源不合规，可能涉侵犯他人知识产权等。而且因其技术特点，数据隐私保护的“知情同意”和数据收集使用“最小必要”原则面临难以落地的伦理风险。以上问题对保障人的隐私和合法权益带来风险，违背了生成式算法“第一定律”。

隐私攻击、模型越狱、数据中毒、基于指令和非指令的后门攻击，是当前较为频繁且重要的针对生成式模型的攻击类型，均会造成个人信息和重要数据泄露[6]。

2.4 算法的可解释性、透明性、可追溯性和技术自主性不足，引发信任和责任难点

生成式算法通常被认为是一种“黑盒”模型，内部工作机制和决策过程难以解释，透明性较差。一方面，生成的内容难以理解和追踪，当模型生成违规内容时，确定责任归属成为难题，目前归责于服务提供方；另一方面，生成式算法在许多应用场景中，难以解释其决策依据和结果[7]。如在一些需要解释决策依据的领域，如医疗诊断、信贷评估、司法判断等，这个问题会很突出，引发信任和责任问题。由于算法决策过程难以解释，对其进行审计以确保其合规性和公平性也变得困难。决策结果难以令人信服，可能导致用户对算法的决策结果产生怀疑和不信任，影响其在实际应用中的接受度和可用性[8]。

尽管生成式算法自2022年以来异军突起，但其技术成熟度和可靠性仍存在一定局限性。Transformer架构的生成式算法可能会产生模式崩塌问题，导致生成内容缺乏多样性和质量稳定性。此外，目前，我国大语言模型算法均基于Transformer架构，使用PyTorch框架，训练算力还主要依赖于英伟达的GPU及其CUDA并行计算架构，很多工程化方法也借鉴国外专业论文成果，自主创新性不足。以上，与生成式算法“第三定律”不符。

3 优化思路与框架

3.1 从数据源头、训练方法、引入评估3个层面纠偏和降低幻觉，加强价值观对齐

降低模型的幻觉、提升算法的公平公正，加强价值观对齐，遵循3个共性思路。

一是对数据去噪纠偏。对训练语料进行去噪和清洗，去除违法违规风险数据、无意义数据、填充缺失值、文本规范化等，消除潜在的偏差、偏见歧视和不符合价值观的数据；对文本进行词频统计，对图像进行分类，找出可能带有偏见歧视的词汇或图像，进行替换或删除；使用数据增强技术提高数据集的多样性。

二是优化算法，强化学习与对抗。更改学习率、使用正则化技术，优化算法。Transformer架构处理长文本有优势、可用来提高对价值观的敏感性和准确性，但生成文本时会出现幻觉，可尝试使用多模型进行融合。使用强化学习，奖励符合价值观的生成结果、惩罚不符合结果来引导模型。使用对抗训练，即在原始数据上训练主模型，在另一个对抗性数据集上训练一个对抗性模型，主模型和对抗性模型迭代优化，减少对特定群体的偏见[9]。这两种方法可提高模型对不良内容的鲁棒性。复旦大学NLP团队在RLHF阶段运用PPO（近端策略优化）算法并优化为PPO-max，让模型更好地理解深层语义。对齐训练后，相较SFT模型，生成内容经测试更符合人类价值观[10]。

三是引入评估指标。针对幻觉问题，可使用困惑度（衡量预测下一个词时不确定性，评估预测效果）、BLEU评分（比较机器翻译与人工翻译间的语法重叠度以评估翻译质量）、ROUGE评分（比较系统生成和人工生成文章间的共现词以评估文摘质量）等指标，评估生成效果。针对歧视偏见，可引入群体公平性指标（比较不同性别、种族等群体在模型决策中的表现，如比较男性和女性申请人在招聘中的录取率差异，如很大，可能存在性别偏见）和反事实公平性指标（比较实际结果和反事实结果之间的差异，如比较一位女性在实际情况下和假设其是男性情况下的决策结果间的差异，如很大，可能存在性别偏见）[11]，以确保模型公平对待不同群体。针对生成内容价值观对齐，可使用安全评估（对标《生成式人工智能服务管理暂行办法》）第四条要求答题测试）、伦理评分（公平性、透明度、责任感等）、语义相似度评估（计算生成内容与预期价值观语义相似度）等方法[12]。

因幻觉问题的解决难度大，在共性方法上，目前还有几类增强方法。一是后处理。使用语义分析、情感分析，识别和修正生成文本中的幻觉问题。二是联网增强或知识库检索。针对一些知识性问题，通过外挂知识库、增加联网组件等检索增强；针对一些涉及国家政权、国家主权等原则问题，建立权威问答库，防止瞎答造成误导或意识形态问题。三是领域适应。使用领域特定数据来微调模型，或使用多任务学习来训练模型，提高其泛化能力。

3.2 降本增效，量力而行，动态平衡参数规模与适用好用之间的关系

发展过程中，追求能力提升与资源耗费、模型可解释性下降的矛盾，可以通过“降本增效”来解决。一是模型压缩与加速。通过知识蒸馏和模型剪枝等压缩和加速技术，在保持性能不变的情况下，降低模型计算复杂度和参数数量，减少资源耗费。二是简化模型与调整训练策略。如Transformer-XL等模型在保持性能的同

时，简化了架构，降低模型复杂度，提高可解释性和训练效率。使用更好的优化算法、更改学习率可帮助模型更有效地学习。强化多任务学习，提高模型泛化能力。三是数据预处理。数据增强、去噪可帮助模型更好地学习数据规律，提高性能。

对于模型研发、运营方来说，要综合考虑占有或可获得的计算资源、训练时间的承受度、训练数据量的收集和预处理能力、算力和时间带来的投资成本，以及模型的应用场景、部署成本等因素，遵循“量力而行”和“适合自己就是就好的”基本原则。

3.3 分类、分级、分场景，加强安全防护和用户隐私保护，强化全供应链安全

针对算法使用、优化、应用中涉及的数据安全和个人信息保护问题，笔者提出：尝试基于数据分类分级的安全防护，提升针对性；尝试不同情形下的“推定同意”“明确同意”“再次同意”，提升知情同意的可操作性；尝试分阶段的不同数据匿名化和加密技术手段，提升有效性；强化事后监管，根据泄露的数量等级，予以分级问责与应急处置。

在此基础上，强化全供应链安全也至关重要，包括模型前后端系统、应用的网络安全，第三方数据提供、标注处理等外包安全和可控性，模型的鲁棒性和抗攻击性等。特别是针对当前模型越狱、后门攻击、推理攻击等攻击方法，通过对模型进行对抗训练、融合多模型等方法，使用差分隐私、安全多方计算等技术，提高模型的鲁棒性，不断加固模型。

3.4 引入伦理规范，优化技术方法，实施分类定制，促进创新发展

伦理规范上，通过道德准则方法（引入伦理规则和约束条件）、价值敏感设计（在算法需求分析、设计、开发、测试、部署全过程，考虑道德、伦理和社会影响）来引导模型决策，在决策过程中加入伦理评估机制，推动决策过程符合人类伦理价值观和道德标准；公开模型的伦理规则、约束条件、评估标准等，使决策过程更透明；记录模型的决策过程、伦理评估结果等，增强可追溯性[13]。此外，开展用户教育也有必要，使各类用户了解生成式算法的原理、局限性和风险，以免陷入“乌托邦”或“敌托邦”的极端。

技术方法上，引入注意力机制、模块化结构、知识图谱等方法，帮助用户理解模型在不同任务中如何工作及决策原因；公开模型架构、训练数据、训练过程等信息，帮助用户更好了解模型内部工作机制；记录模型训练过程、参数更新、数据来源等，帮助用户进行故障排查和问题定位。

分类定制指针对不同群体、不同应用场景，开展专用的算法定制和优化。了解不同群体、不同场景的需求，与心理学、社会学专家和领域专家合作，针对性地设计和优化算法，并加强用户教育。如针对未成年人，设计算法要考虑加强内容过滤、时间管理和教育支持；针对老年人，提升易用性、帮助健康监测和咨询、帮助其与家人朋友联系社交；针对消费者，算法优化价格比较、评价分析、售前售后服务等保障；针对女性，要保障性别公正、提供女性健康建议、职业发展支持等。在重要领域，行业主管部门牵头制定实施合规和伦理指引。医疗领域，算法要保护患者隐私、提升诊断和治疗建议的准确性，以免误导医生或患者，可使用特征重要性分析方法，帮助理解不同特征对模型预测结果的贡献程度，如疾病预测场景中，可找出对疾病预测最重要的特征，为医生诊

断提供参考。金融领域，提高算法合规性，在信贷、保险等产品定价审批中的公平性。可使用反事实解释方法帮助理解算法决策过程，如在信贷审批场景，告诉申请人为什么贷款申请被拒绝，由哪些因素导致。司法领域，要提升算法合法合规性，提高辅助法官决策的公正性，减少误判概率以确保司法公正。教育领域，算法要保护学生隐私、服务不同学生的学习需求和进度、保障资源分配公平，还要避免过度依赖，以免影响学生自主学习能力和发展。以上各场景，算法都要提高可解释性，以便不同用户（医生和患者，用户与监管部门等）理解算法决策过程。

在全球激烈竞争的格局下，生成式人工智能领域不发展就是最大的不安全。因此，目前，我国对生成式人工智能服务采取“包容审慎”监管原则，对技术自研自用基本没有约束限制；鼓励算法、框架、芯片及配套软件平台等基础技术的自主创新，参与国际规则标准制定。

4 可能仍存在的难点和困境

一是机器幻觉问题无法根治。按目前语言大模型业界共识，即使采取强化数据清洗、改进模型架构和训练策略、引入检索增强和事实校验等各类优化方法，仍只能将生成内容的准确性、可靠性最高提升到约80%。剩下的20%，是现阶段技术的盲区。

二是评估审查规则、量化指标存在局限性。不同利益相关方在算法伦理评估和审查中可能持有不同价值观和道德观，可能导致各方在评估审查规则、指标等方面难以达成共识。前文所提各类评估指标多数是评价算法模型性能。因商业驱动，业内已盛行通过针对性“刷榜”来提升自家模型“考试成绩”。但是，在生成内容的安全性、价值观符合性方面，目前并没有成熟的量化评估或审查机制，特别是伦理问题的复杂性，可能很难用指标来评估。

三是评估审查与算法迭代速度间的矛盾与平衡难点。生成式算法模型的伦理审查和安全评估涉及制定审查评估规则、多方参与、确定指标、将评估审查纳入整个生命周期，持续优化、反馈循环等流程，除了技术、方法的难点外，可能带来时间、人力、资金等资源限制和投入，与生成式厂商们以OpenAI为目标、持续搞算法模型“炼丹”迭代升级之间存在矛盾。要真正做到发展和安全的动态平衡、相得益彰，可能是一个长期复杂的过程。

因此，需要通过不断完善生成式人工智能的监管机制和伦理框架，推动产业链相关主体共同发挥作用，随着技术的不断发展，持续探索与完善。

5 结束语

本文聚焦生成式人工智能的生成式算法，结合相关法律法规、伦理规范，提出“生成式算法三定律”（三大定律，12条指引）的伦理原则。同时，结合其技术特点，分析算法伦理在实践中存在的四大难点，并初步提出一些解决和优化的框架和思路。然而，因各类原因，这些优化框架还存在一些伦理困境，需要多方合力，长期探索、完善。

参考文献：

[1] 习近平. 全球人工智能治理倡议[R]. 第三届“一带一路”国际合作高峰论坛, 2023.
XI J P. Global artificial intelligence gov -

ernance initiative[R]. The 3rd "The Belt and Road" International Cooperation Summit Forum, 2023.

[2] 万小军. 智能文本生成: 进展与挑战[J]. 大数据, 2023, 9(2): 99-109.
WAN X J. Intelligent text generation: recent advances and challenges[J]. Big Data Research, 2023, 9(2): 99-109.

[3] DZIRI N, MILTON S, YU M, et al. On the origin of hallucinations in conversational models: is it the datasets or the models? [EB]. arXiv preprint, 2022, arXiv: 2204.07931

[4] HEATON J. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: deep learning[J]. Genetic Programming and Evolvable Machines, 2018, 19(1): 305-307.

[5] 施敏, 杨海军. 大语言模型的数据隐私保护难点分析与探索[J]. 大数据, 2024, 10(5): 168-176.
SHI M, YANG H J. Difficulties and explorations in data privacy protection for large language models[J]. Big Data Research, 2024, 10(5): 168-176.

[6] 参赞生命力. 浅谈大模型数据隐私[Z]. 绿洲资本, 2023.
CAN Z. An introduction to big model data privacy[Z]. Vitalbridge, 2023.

[7] SAMEK W, MÜLLER K R. Towards explainable artificial intelligence[M]//Explainable AI: interpreting, explaining and visualizing deep learning. Cham: Springer, 2019: 5-22.

[8] PFEFFER A. Trustworthy machine learning: mitigating the risk of unintended consequences [EB]. arXiv preprint, 2021, arXiv: 1901.10002v3.

[9] BAROCAS S, HARDT M, NARAYANAN A. Fairness and machine learning[M]. [S.l.: s.n.], 2019.

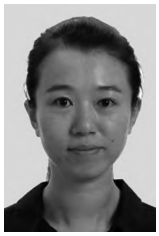
[10] ZHENG R, DOU S H, GAO S Y, et al. Secrets of RLHF in large language models Part I: PPO. 2023.

[11] PESSACH D, SHMUELI E. A review on fairness in machine learning[J]. ACM Computing Surveys, 2022, 55(3): 1-44.

[12] GÉRON A. Hands-on machine learning with Scikit-Learn and TensorFlow[M]. [S.l.: s.n.], 2019.

[13] GELFERT M, MITCHELL M, LEE K. Towards a conceptual framework for ethical considerations in natural language processing[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.]. 2020.

作者简介



施敏（1981-），女，就职于上海市互联网信息办公室，主要研究方向为互联网新技术新业务的发展和安全、网络空间治理、人工智能与大数据伦理等。

杨海军（1973-），男，博士，上海市互联网信息办公室高级工程师，主要研究方向为通信和公共互联网领域的网络与信息安全工作、互联网舆情、网络空间治理、网络与信息安全战略、互联网新技术新业务的发展及其安全等。

收稿日期: 2024-02-02