·理论研究·

大模型伦理失范的理论解构与治理创新

肖红军[1]，张丽丽[2]

（1. 中国社会科学院 工业经济研究所，北京 100006；
2. 北京市科学技术研究院 数字经济创新研究所，北京 100089）

摘　要：ChatGPT系列和Sora等大模型作为人工智能技术的发展前沿，在助力人类生产生活的数字化和智能化转型提速的同时，也引发一系列伦理失范现象，成为全球性治理难题。本文对大模型伦理失范进行理论解构发现：大模型伦理失范的缘起是其区别于传统生产力和生产工具的技术非中性、内嵌人类伦理和将人类伦理与机器伦理糅合其中等属性特征，技术起点是算法，关键中介是数据，行为主体是人类。大模型所引发的诸如模型黑箱、数据版权侵害且主体责任难确定、冲击人类主体资格等伦理失范现象对现行伦理治理模式提出挑战。因此，本文以大模型生命周期为时间维度，以关键要素为核心，构建了基于大模型关键要素的全生命周期伦理治理框架。同时，为推动新伦理治理框架的有效运行，本文构建了包含两层行为主体的自我治理、两级守门人和全球性合作治理网络等子体系的大模型伦理治理生态体系。

关键词：大模型；伦理失范；伦理治理

中图分类号：F49；B82-057　　**文献标识码：**A　　**文章编号：**1000-176X（2024）05-0015-18

一、引　言

人工智能技术创新日新月异，不断实现新突破。以OpenAI的ChatGPT系列和Sora、谷歌的Gemini、百度ERNIE 3.0、华为盘古等为代表的语言大模型和视频大模型是人工智能技术的发展前沿，其在海量数据、强算力和算法三个关键因素共同推动下，模拟人类思维链运行。自2022年底ChatGPT问世以来，大模型领域逐渐形成国际上以OpenAI和谷歌为代表，国内以百度、华为和阿里巴巴等为代表的技术研发和应用终端布局体系。其中，2024年2月，OpenAI推出的Sora将大模型技术从文字式智能涌现跃迁至视频式智能涌现，从静态式呈现跃迁至动态式呈现，Sora成为链接物理世界和数字世界的“通用模拟器”。与其他人工智能应用相比，大模型实现了突破性技术创新和颠覆性技术创新[1]，且依靠其通用性、多模态和智能涌现能力与千行百业深度融合，引发生产方式、技术创新范式、内容生成方式和人机关系等领域的变革。例如，科学研究、

收稿日期：2024-03-22
基金项目：国家社会科学基金重大项目“国企混合所有制改革的实现路径选择研究”（20&ZD073）；中国社会科学院国情调研重大项目“数字科技伦理监管制度调查研究”（GQZD2023009）；中国社会科学院登峰战略企业管理优势学科建设项目
作者简介：肖红军（1977-），男，湖南郴州人，研究员，博士，博士生导师，主要从事企业社会责任与数字治理研究。E-mail：xiaohjxiaohj@126.com
张丽丽（通讯作者）（1985-），女，山东临沂人，副研究员，博士，主要从事产业经济与数字治理研究。E-mail：526lily@163.com

生物医药和软件开发等领域从分散个体化的“慢工出细活”式生产演变为平台集成化的“短平快”式智能涌现。然而，大模型应用也会引发一系列伦理失范现象。例如，2023年3月31日，ChatGPT就因为用户隐私泄露等问题在意大利被禁止使用，并因涉嫌违反数据收集规则被调查。2024年1月29日，意大利数据保护局称OpenAI违反欧盟《一般数据保护条例》，可能会对其处以全球营业额4%的罚款。此外，大模型伦理失范现象还包括算法黑箱、数字鸿沟和隐私泄露等，并衍生出新型伦理失范现象。例如，模型黑箱、数据版权侵害、深度伪造加深、冲击人类主体资格和加剧社会阶层分化等。

如何有效防范与治理大模型等生成式人工智能应用所引发的伦理失范成为全球性重要议题，“科技向善”“负责任研究与创新”“科技伦理”等理念被提出并付诸实践，国际组织、各国家或地区纷纷出台相关伦理监管政策以实现伦理失范治理与技术创新之间的平衡[2]。其中，国际组织中具有代表性的监管政策包括，联合国教科文组织发布的《人工智能伦理问题建议书》、G20发布的《G20人工智能原则》、电气与电子工程师协会发布的《人工智能设计伦理准则》、G7发布的《开发先进人工智能系统组织的国际行为准则》、欧盟发布的《人工智能法案》和配套的《人工智能责任指令》、28个国家和欧盟共同签署的《布莱切利宣言》，以及英国、美国等18个国家联合发布的《安全人工智能系统开发指南》等，从正式制度和非正式制度两个层面加强人工智能伦理治理。具体到单个国家或地区层面，美国的伦理监管模式逐渐从相对宽松转向趋紧监管，且以总统拜登签署的《关于安全、可靠和可信的人工智能》行政命令为分水岭；中国高度重视大模型相关伦理治理，出台了《新一代人工智能伦理规范》《科技伦理审查办法（试行）》《生成式人工智能服务管理暂行办法》等一系列政策文件，逐渐加强伦理治理的制度建设，并新设国家科技伦理委员会以推进科技伦理治理。其中，《生成式人工智能服务管理暂行办法》提出，对生成式人工智能服务实行包容审慎和分类分级监管，且明确服务提供者和应用者等主体的责任。为提高大模型伦理治理的法治化水平，《国务院2023年度立法工作计划》将人工智能和网络数据安全管理相关立法提上日程。英国、新加坡等国家也出台了相应的监管政策以防范和治理大模型等生成式人工智能的伦理失范。另外，为加强大模型伦理治理，各国家或地区还设置了专门监管机构。例如，欧盟《人工智能法案》提出，设立欧盟人工智能办公室（European AI Office）以提高通用型人工智能模型的透明性，以及控制系统性风险等。同样，日本提出要成立一个全国性的机构对人工智能伦理失范进行研究与监管。然而，现行的单要素和单环节、自上而下和以结果为导向的伦理治理模式无法有效防范和治理大模型等生成式人工智能应用所引发的算法偏见、隐私泄露、数据版权侵害等伦理失范现象。为加强大模型伦理治理，亟须对大模型伦理失范的缘起进行探究，对伦理失范现象进行归纳，对相关伦理治理范式进行梳理，进而针对现行理论研究薄弱之处、治理制度体系待完善之域和治理实践待创新之区等，探寻更加适宜和有效的伦理治理范式。

鉴于此，本文对大模型的属性和伦理失范缘起进行分析，解构大模型伦理失范的关键要素，归纳大模型伦理失范的主要表现及其对现行治理体系提出的挑战。在此基础上，构建大模型伦理治理的新框架和相应治理生态体系。从理论上对大模型伦理失范缘起给予回答，从治理框架上进行创新，从治理生态上进行构建，以期实现大模型伦理治理秩序重构的目标。

本文的边际贡献主要包括：一是对大模型的属性和关键要素进行解构，是对大模型等生成式人工智能本质的再认识。二是对大模型伦理失范的缘起进行理论解构，进一步揭示算法、数据和人类等关键要素在大模型伦理失范中所发挥的作用。三是对大模型伦理失范现象进行归纳和总结，在人工智能伦理失范现象或风险的基础上进行了拓展。四是以大模型生命周期为时间维度，以关键要素为核心，构建基于大模型关键要素的全生命周期伦理治理框架，并构建相应的大模型伦理治理生态体系，以有效防范与治理大模型伦理失范现象。

二、文献回顾

人工智能伦理相关研究最早可以追溯至1940—1950年关于伦理和数字技术的探讨，此后，紧密贴合科学技术发展脉络，伦理治理相关研究脉络的演变从科技伦理到人工智能伦理，再到大模型等生成式人工智能伦理治理。大模型伦理治理的相关研究主要涌现于2022年底ChatGPT问世之后。与本文相关的文献主要集中于以下三类：一是大模型对社会、经济的影响及其伦理失范的相关研究。二是大模型伦理治理的相关研究，主要集中于治理模式、治理目标和实现路径等。三是大模型关键要素伦理治理的相关研究，以算法、数据和深度合成技术等为主。

（一）大模型对社会、经济的影响及其伦理失范的相关研究

随着人工智能技术的演进，大模型等生成式人工智能对经济增长、生产效率和要素分配等方面产生深刻影响[3]，起到提质增效和深化供给侧结构性改革等作用，与此同时也引发隐私泄露、数字鸿沟、输出带有歧视性或偏见性结果、知识产权归属不清晰等伦理问题，可归纳为技术内生型伦理风险和技术应用型伦理风险两大类[4]。具体到不同领域，大模型会引发不同伦理失范现象。例如，大模型广泛应用于医学、经济和金融等研究领域会引发科研范式变革，能够自主进行科研假设、科学实验和验证假设合理性等全流程[5]，从低效率的作坊模式转变为平台模式[6]，对创新宽度产生影响[7]。此外，大模型还可以为科研人员提供文献回顾、数据访问[8]等辅助性研究便利，但也会导致数据版权侵害且主体责任难确定[9]、知识产权侵害和科研不端等伦理失范现象。其中，主体责任难确定将进一步引发社会信任体系的混乱。大模型还适用于工业设计、药物研发和材料科学等领域，其对简单重复性劳动所产生的替代作用将引发社会分化或歧视等伦理失范现象。

（二）大模型伦理治理的相关研究

为应对大模型伦理失范及其对现行伦理治理体系提出的新挑战，学术界对大模型伦理治理开展了相关研究。一是对大模型进行道德伦理设计，从技术上降低伦理失范发生的概率。基于前期所习得知识和输入信息，大模型通过所构建的思维框架生成内容，具有思维价值[10]并体现伦理内涵。对大模型伦理进行有效治理，可以通过在研发阶段引入伦理专家团队[11]、构建伦理数据库或伦理知识库、价值观嵌入、敏感价值设计、提示工程[12]和内置纠偏机制等方式，使大模型等生成式人工智能习得人类的道德伦理知识体系，以降低伦理失范现象发生的概率。二是大模型伦理治理体系或框架的相关研究。关于大模型伦理治理体系或框架的直接研究对象为生成式人工智能、通用人工智能或通用模型[13]等。目前，较成熟的人工智能伦理治理模式包括规制型治理、创新型治理、自治型治理、市场导向型治理、集中式治理、敏捷治理[14]、复合型系统性治理和面向产业链的韧性治理[15]等。张凌寒[6]提出，针对生成式人工智能要构建“基础模型—专业模型—服务应用”的分层治理体系，形成复合型系统性治理。王沛然[16]提出，针对大模型的特征应遵从控制主义转向训导主义的治理范式，其包含大模型的开发训练动机、训练数据的内容生态建设、具体场景和用户的规制等三个基本法学命题。三是对大模型伦理治理的目标、实现路径和不同主体的法律定位等进行研究。结合《生成式人工智能服务管理暂行办法》等监管政策对大模型服务提供者侵权责任的界定，进一步从生成和移除两个维度进行判定，强调责任判断标准的适时可调整性[17]。考虑到现有的伦理治理政策存在可操作性不强、刚性约束不足和效力层级较低等问题，应坚持法治先行，加快构建科技伦理规范体系[18]。当然，也有学者提出，大模型技术处于尚未完全落地的阶段，不适宜在早期就开展专门的伦理治理立法工作，过早的干预或治理将影响技术创新，可能导致法律偏离原本的功能定位，应该从技术标准、安全风险评估和信息审核技术等方面进行治理[19]。

（三）大模型关键要素伦理治理的相关研究

大模型关键要素伦理治理的相关研究主要集中于数据、算法等。一是关于大模型训练数据伦理治理的相关研究。大模型训练需要海量数据作支撑，数据规模越大，大模型能够学习和提取的人类道德伦理知识就越多[20]，但存在训练数据的误导性、偏见和歧视等伦理风险[21]。为加强数据伦理治理，Someh等[22]提出，要构建个人、组织和社会互动的数据伦理治理框架，包括正式治理和非正式治理，从而合乎道德地使用大数据。续继和王于鹤[23]提出，要构建以综合推进数据要素市场的高效、公平和安全为目标的“个人—企业—社会”的三维共建数据治理体系框架。张欣[24]针对数据的质量、安全和时效等问题，提出构建精准多元的数据主体责任矩阵，打造灵活高效的数据治理监管工具体系。二是对算法伦理治理的相关研究。针对算法黑箱、算法偏见和算法歧视等伦理问题，有研究提出分场景的算法规制方式，构建算法公开、数据赋权和反算法歧视等算法规制的具体制度，以实现算法负责任的目标[25]；针对算法伦理风险问题，有研究提出构建自动化、生态化和全流程的动态监管体系，以实现敏捷治理[26]。肖红军[27]从人、算法和社会三个治理要素的九个维度构建了算法责任综合治理范式的九宫格模型。然而，不同主体算法责任的分配原则和用户的算法责任等问题仍待进一步廓清[28]。此外，张凌寒[29]从深度合成技术伦理治理的视角切入，提出对大模型实施全链条治理，在全球形成更具影响力的治理法律制度体系。

通过以上分析可以发现，现有研究仍然存在以下不足之处：一是数据、算法等关键要素伦理治理的研究自成独立体系，对大模型关键要素的伦理治理缺乏针对性和适用性。二是未对大模型生命周期的不同阶段和关键要素进行拆解，未对关键要素的伦理失范缘起进行系统分析，目前也未有相应的伦理治理框架。大模型的技术复杂性、结果的不确定性和应用的广泛性所引发的伦理风险呈现出影响范围广、传播速度快和乘数效应强等特征，现有的伦理治理模式已经无法应对，亟须提出新的伦理治理框架。

三、大模型伦理失范的理论解构

相较于其他人工智能技术和应用，大模型存在伦理失范现象的根源是什么，要从分析其本质属性开始。本文运用马克思关于生产力和生产工具的相关理论对大模型的属性进行解构，并结合大模型生命周期的主要阶段，对大模型属性及伦理失范缘起进行探究，为进一步防范和治理大模型伦理失范现象寻求路径。

（一）大模型属性及伦理失范缘起

1.生产力属性：技术非中性

大模型是人工智能技术的发展前沿，是深度神经网络（DNN）在强算力基础设施和海量数据支撑下，算法及框架的叠加与演进而成的集成性技术系统。在Transformer架构的基础上，神经网络模型逐渐演进为三类大模型典型架构，即掩码语言模型、自回归语言模型和序列到序列模型。作为一项人工智能技术，大模型隶属于生产力的范畴，且属于以科技创新为核心的新质生产力，是推动社会进步最活跃的要素[30]，是社会生产力的一次跃升[31]。进一步深入分析，马克思从两个维度对生产力进行阐释和分类：生产力是人类进行物质资料生产的客观力量，即物质生产力；生产力是人类生产精神财富的能力，即精神生产力。两者的统一性体现为物质生产力和精神生产力均形成了使用价值。从这一意义看，大模型的生成内容以满足人类的精神需求为主，其属于精神生产力的范畴。

大模型的核心技术要素为算法，其也属于生产力的范畴，具有黑箱特征。算法是从数据输入至结果输出的一项计算机计算规则或步骤，即算法=逻辑+控制[32]。算法黑箱可划分为主观算法黑箱和客观算法黑箱[33]，并进一步衍生出算法偏见、算法歧视等伦理失范现象。算法黑箱产生

的缘由可归纳为三点：一是从知识产权的角度看，算法具有商业秘密性和经济价值性，无法向公众公开。例如，大模型GPT-4是可以复制的，但OpenAI不会将其公之于众。二是算法作为技术，其本身具有不透明性和不可解释性，人类无法预知其产生的结果。三是算法黑箱体现为不同人群对算法的运算机制和决策程序等信息存在认知偏差，这种偏差尤其存在于专业人员与非专业人员之间。算法黑箱在大模型伦理风险中体现为应用者无法观察或验证算法的数据处理过程，导致结果不确定性和不可解释性等伦理风险。因此，从生产力范畴的角度分析，大模型不再具有“价值中立性”，存在算法黑箱及其衍生的伦理失范现象。

从所作用的劳动对象来看，大模型内嵌人类的伦理倾向或偏好，不再保持“价值中立性”。与其他生产力相比，大模型所作用的劳动对象是数据，其具有虚拟性、可重复应用且包含人类的道德伦理内容等特征。一是数据作为一种新型生产要素，其存储或管理需要依托于网络等载体，与传统要素相比具有虚拟化的特征。二是一旦数据脱离所有者，数据可以被重复使用和处理，并且控制其未来流向和使用途径等具有一定困难，即存在数据伦理风险敞口。三是数据要素包含人类的价值观等伦理倾向或偏好，尤其是个人数据，具有“人格化”特征。正是由于大模型所作用的数据要素包含人类的伦理倾向或偏好，且自身存在伦理风险敞口，数据参与大模型的运行，并内嵌到生成内容中，伦理偏好或倾向、伦理风险也将进一步传递。

2.生产工具属性:内嵌人类伦理

大模型作为一种应用系统，是人类作用于输入内容的中介，使输入内容发生一定变化，是人脑的延伸，因而大模型属于生产工具的范畴。人类在大模型的研发过程中融入了自身的伦理倾向或偏好；在大模型的使用过程中产生包含伦理倾向的内容，并对人类产生伦理影响。

在大模型的研发阶段，研发者的伦理倾向或偏好注入其中，因而算法黑箱不可避免。人工智能技术不断演进，大模型在尽力模拟人类大脑的运行方式，参数规模力求接近于大脑神经突触数量。例如，大模型GPT-4由8个混合模型组成，每个模型约2 200亿个参数，总共约1.8万亿个参数，是大模型GPT-3的10倍以上。其中，算法模型的选择和参数规模的确定均取决于研发者的主观意愿。因此，随着大模型研发技术的不断演进，大模型的运行模式将越来越接近人类大脑，研发者及所模拟人类思维中所包含的伦理倾向或偏好会渗透其中，算法的黑箱特征在大模型整个生命周期中如影随形。

在大模型的训练阶段，训练数据、训练人员和数据处理者的伦理倾向或偏好会融入其中。一是训练数据的来源和质量本身包含伦理倾向或偏好。按照是否经过处理或加工，可以将训练数据划分为原始数据和合成数据。大模型的训练数据大多来源于公开数据库，未进行加工或合成。由于来源确定或属于公开数据库，原始数据应重视其内容的正当性。为提高大模型训练数据质量，降低对训练数据规模的依赖性，需要对训练数据进行质量过滤、冗余去除、隐私消除等处理，在该过程中就会将人类的伦理倾向或偏好内嵌其中，且可能因训练数据规模的庞大和内容的复杂性等因素而无法进行精确处理，即存在伦理风险敞口。合成数据与原始数据的根本区别在于除数据本身所包含的伦理倾向或偏好之外，数据合成过程中掺入了数据处理者的伦理倾向或偏好，影响其内容的正当性。二是训练数据的处理和标注方法等包含着人类的伦理倾向或偏好。大模型预训练数据以原始数据为主，合成数据为辅，调优数据以标注数据为主。大模型调优数据主要包括代码类数据和对话类数据。其中，代码类数据主要来源于GitHub上的公共代码库，且以Python文件为主；对话类数据以人工标注数据为主，数据标注者主要包括专业标注工和大模型应用者。因此，大模型训练数据会受数据处理者、数据标注者和应用者等群体的伦理倾向或偏好的影响，且存在“漏网之鱼”的伦理风险敞口和“以偏概全”的局限性。

大模型作为生产工具，不仅是人类体力劳动的延伸和替代，还是人类大脑和神经系统的延伸和替代。一是大模型是人类体力劳动的延伸和替代，其带来生产工具供给的范式变革。人类具有

先天生理上的限制。例如，人类无法在极高或极低温度下进行生产劳动。为克服人类自身的生理性局限，生产工具作为劳动者和劳动对象的中介应运而生。生产工具是人类作用于劳动对象的物件，是人类社会时代划分的重要标志。机器是工业社会的重要体现，而以人工智能技术支持的机器人等智能化机器是数字经济时代的重要体现。区别于其他生产工具，大模型在应用过程中主要作用于数据要素，其生成内容也具有一定的虚拟性和数字化特征，拥有智能涌现且与人类思维相似等特征，带来了生产工具供给的范式变革。二是大模型是人类大脑和神经系统的延伸和替代，大模型应用过程中将人类的伦理倾向或偏好融入其中。大模型作为智能化生产工具引发了“代具性”“技术附庸”问题，严重冲击人类主体资格。机器大生产代替传统手工生产，人类的劳动不断趋向重复性和机械性，加剧了人的物化和异化[34]，即在以机器为中心的生产体系中，人被嵌入机械化生产线中，劳动强度被不断提升，引发精神和体力的双重压力，制约了人类的全面发展。与此类似，数字经济时代，大模型高度渗透到人类生产生活中，人与机器不断互嵌互构，从机器服务于人、人机协作、人机共生，进一步演变为人机一体，人与机器的边界不断模糊，久而久之人类会对大模型产生过度依赖，形成“代具性”，甚至沦为“技术附庸”[34]。这会引起人的主体性消解和决策自主权丧失[35]等现象，严重冲击人类主体资格。三是大模型在训练过程中会学习训练数据的特征、结构、知识和伦理内容等，并延伸至其所生成内容的伦理倾向，对人类社会产生伦理影响。

3.双重属性：人类伦理与机器伦理糅合

大模型是算法和模型等人工智能技术的集成系统，与应用端紧密联系在一起，是一种新型生产力工具[36]，又被称为数字化生产力工具。以OpenAI的ChatGPT系列和Sora为例，其底层技术是Transformer架构和Diffusion Transformer架构，应用端是大模型GPT-4和Sora。大模型是生产力与生产工具的集合体，兼具两者的属性特征，从研发阶段到应用阶段均将人类的伦理倾向或偏好内嵌其中，且生成带有伦理倾向或偏好的内容。波普尔认为，科学实践的特点是不断地根据经验检验理论，并根据检验的结果进行修正。由该理论可知，大模型是科学技术与科学实践的融合，是人类思想或意识在机器上的延伸。与马斯克的脑机接口直接读取人类思想不同，大模型通过对人类思想或意识进行学习之后形成涌现能力，是人机一体化不断演进的一种表现，更是机器与人类伦理道德体系不断糅合之后的外在展现。因此，大模型作为一种数字化生产力工具，兼具生产力与生产工具的双重属性，伦理失范的缘起或底层逻辑是两者的综合，并贯穿于大模型生命周期（如表1所示），伦理失范现象更加复杂且难治理。

表1　大模型生命周期及伦理失范缘起解析

失范缘起	研发阶段	训练阶段	应用阶段
失范因素	算法、模型	预训练数据、调优数据	用户数据、生成内容
失范主体	研发者	数据处理者、数据标注者和部分应用者	服务提供者、应用者
失范现象	算法黑箱、模型黑箱等	数据安全、隐私侵犯和数据版权侵害等	算法偏见、深度伪造、数字鸿沟、社会阶层分化、数据安全、隐私侵犯、数据版权侵害和冲击人类主体资格等

注：表中内容来源于公开信息，由作者整理得到。

（二）大模型伦理失范的关键要素解构

大模型是人工智能技术的发展前沿，通过生产力、生产工具及二者集合体这三个层面进行解构可以发现，算法、数据和人类是引起大模型伦理失范的关键要素。算法是大模型伦理失范的技术起点，数据是大模型伦理失范的关键中介，人类则是大模型伦理失范的行为主体。大模型生命周期、关键要素与伦理失范现象的关系如图1所示。

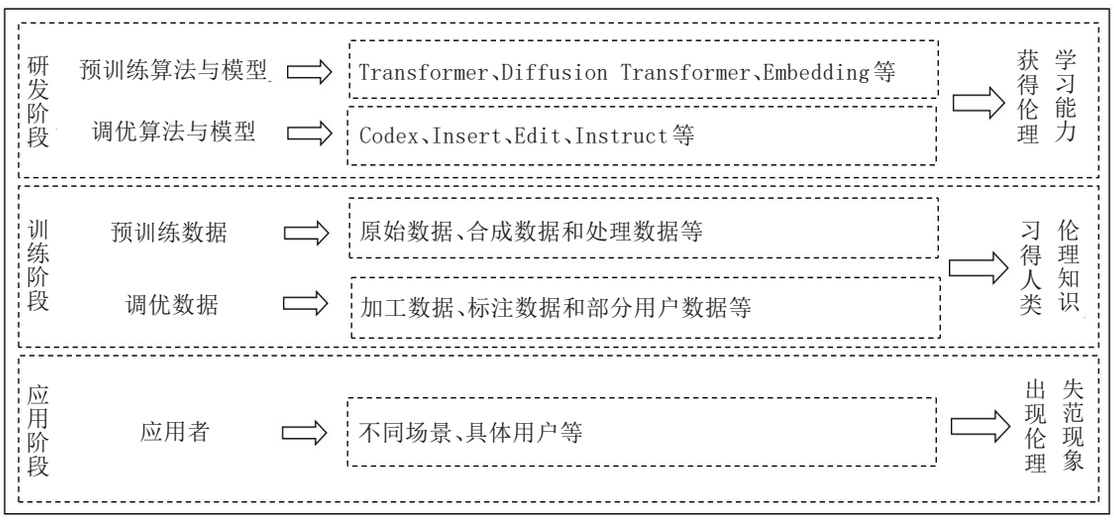

图1 大模型生命周期、关键要素与伦理失范现象的关系

1.大模型伦理失范的技术起点:算法

算法是大模型运行的内在逻辑，依靠运行规则和学习能力对“投喂”其中的训练数据进行学习，生成人类满意的内容，其“黑箱”问题也随之延伸至生成内容，并体现为生成内容的不可追溯性和不确定性。另外，算法通过对数据进行学习，并部署于不同的应用场景中，嵌入人类的社会生活中，其“价值中立性”不复存在，成为局部人类道德伦理体系的习得者、反映者和影响者。如果算法黑箱变为算法透明箱，其内在运行逻辑显而易见，大模型的生成内容就可预测，伦理失范的缘由则可追溯、可修正。因此，算法是大模型伦理失范的技术起点，也是算法黑箱、算法歧视和算法偏见，以及所衍生的模型黑箱、模型歧视和模型偏见等伦理失范现象的根源。

2.大模型伦理失范的关键中介:数据

大模型的出现将传统的知识生成模式转变为机器生成模式，且以类人脑的方式在短期内实现生成内容的智能涌现。大模型的训练和运行依赖于数据，且学习和传承了内嵌其中的人类道德伦理体系，通过“镜像效应”[13]使其生成内容表征局部人类道德伦理体系，并对人类道德伦理体系产生影响。按照大模型数据“投喂”目的的不同和所形成的道德伦理体系的先后顺序，将伦理道德体系划分为原生道德伦理体系和衍生道德伦理体系。训练数据的“投喂”形成大模型的原生道德伦理体系，是研发者、训练者、数据标注者等相关主体，以及训练数据中所包含的道德伦理体系在进行模型研发、训练和优化过程中形成的初始道德伦理体系。衍生道德伦理体系指大模型进入应用阶段，应用者将自身道德伦理或输入内容所包含的道德伦理等输入大模型，在智能交互中对原生道德伦理体系产生一定影响，并在其基础上所形成的新道德伦理体系或引起原生道德伦理体系的变化。大模型的道德伦理体系将研发者、训练者、数据标注者及其输入内容的道德伦理体系进行糅合。其中，输入内容作为客观主体体现了前述相关主体的道德伦理，因而大模型所习得的道德伦理体系是人类道德伦理体系经过算法等人工智能技术进行处理之后形成的多个人类主体的道德伦理融合体系。

3.大模型伦理失范的行为主体:人类

大模型的研发者、训练者、服务提供者和应用者是人类，训练数据的产生者、处理者和输入者也是人类，因而引起大模型伦理失范的行为主体是人类，受影响的主体是人类，承担责任的主体也是人类自身。算法是人类的代理者[27, 37]，大模型同样充当着人类代理者的角色，区别在于所代理内容多少和程度大小。数据是人类的产物和资源，是人类道德伦理体系的反映物，从这个意义上讲，数据具有“价值中立性”的特征，是否产生伦理失范现象取决于人类的生产方式、应

用方式和具体场景。综上，尽管人类的主体资格受到大模型的冲击，但人类依然是大模型的控制者和其生成内容的决策者。大模型体现的是人类的意志，大模型等生成式人工智能与人类不论承担交互式伦理责任，或完全式伦理责任[38]，或无需承担伦理责任[39]，大模型伦理失范行为主体、受影响主体和责任主体最终是人类自身。

（三）大模型伦理失范的主要表现

对大模型伦理失范的缘起进行分析与解构后发现，大模型具有算法叠加、参数规模增加、训练数据来源多元化和规模庞大、智能涌现能力更强等特征，这引起了新型伦理失范现象的发生和原有伦理失范现象的深化或扩散。例如，模型黑箱、数据版权侵害且主体责任难确定、冲击人类主体资格和对人类社会道德伦理秩序的漠视和扰乱等。

第一，模型黑箱。在关于人工智能伦理、算法伦理的分析和研究中，算法黑箱是伦理治理的焦点和难点，包括由此衍生出的算法歧视和算法偏见、结果不确定性和误导性[40]等伦理失范现象。在算法黑箱的基础上，依赖深度神经网络和集成技术的大模型所产生的黑箱现象演变为模型黑箱[41]。算法是从数据输入至结果输出的一项计算机计算规则或步骤，即算法=逻辑+控制[32]，深入渗透到社会生活中的算法已经成为社会价值判断的一部分，不再具有“价值中立性”[25]；参数是模型中需要训练和优化的变量，将影响模型的生成内容，参数越多，模型的生成内容越符合预期；模型是算法使用数据进行训练后输出的文件，是具有特定流程或结构的计算机程序[16]，算法和参数均包含其中。模型黑箱比算法黑箱更加复杂，是算法黑箱的叠加，是算法、模型和参数等要素糅合之后的“黑箱”。模型黑箱进一步加剧了大模型生成内容的不可解释性和不确定性，甚至其生成内容是错误的、虚假的、有害的、具有误导性且缺乏因果推断的[42]，加之存在“机器幻觉”现象，会对人类行为产生不良诱导。与专用人工智能模型相较而言，大模型的智能涌现能力强，内容生成效率高，易于在短期内形成劣质内容且能快速传播，在一对多的连锁反应下将对人类社会道德伦理秩序带来系统性冲击。

第二，数据版权侵害且主体责任难确定。数据和文本挖掘技术为公开数据的获取与使用提供了便利。大模型的智能涌现能力依赖于海量数据的“投喂”，然而大模型所使用数据的获取途径和使用方式是否正当、合法仍需进一步加以监管。其中，欧盟理事会通过的《数字化单一市场版权指令》对获取、使用数据和文本进行了规范，提出仅不受版权保护的事实或数据无需授权，即数据获取和使用的主体授权原则，但仍需注意是否存在复制权侵害。大模型的部分训练数据直接来源于网络爬取，且未对数据的合法性和安全性等进行评价和验证，如果爬取数据超出了数据主体的授权范围或包含敏感信息等，将会侵害数据主体权利，又由于人机边界模糊、数据来源模糊和数据主体权利模糊等延伸出主体责任难确定等伦理问题。大模型训练使用合成数据同样存在上述问题。其中，Sora视频大模型的出现将数据版权侵害问题扩展至视频生成领域。

第三，人机边界模糊，物理世界与虚拟世界界限模糊，这将会冲击人类主体资格和社会信任体系。政治经济学关于生产关系的相关理论认为，人与生产工具、人与生产力之间均是彼此独立的二元结构。数字经济时代，大模型与人的关系打破了上述二元结构，加剧了人的物化和异化[34]。从语言大模型ChatGPT到视频大模型Sora，大模型技术的演进体现为人与机器不断互嵌互构，人机边界不断模糊，引起人的主体性消解和决策自主权丧失[35]等现象，严重冲击人类主体资格。此外，视频大模型Sora的问世使深度伪造进一步加深。深度伪造是基于生成对抗网络模型，广泛应用于视频换脸的人工智能技术，凭借其高度真实性、泛在普适性和快速演化性等特征，其应用不当会产生侵犯公民人身财产权利、消解社会信任体系等伦理问题。与深度伪造技术相比，视频大模型取得了实质性的超越，加深了伪造的深度，从对人脸的“伪造”演变为对物理世界的“伪造”，进而影响人类社会信任体系。

第四，对人类社会道德伦理秩序的漠视和扰乱。一是大模型的生成内容对人类道德伦理秩序

的漠视和扰乱。具体到应用场景的大模型，携带着所习得的局部人类道德伦理体系，生成应用者所需要的内容，并将原生道德伦理体系和衍生道德伦理体系糅合其中。大模型的研发者、服务提供者和应用者漠视生成内容是否存在对人类社会道德伦理秩序的扰乱，大模型生成内容处于人类道德伦理体系的真空区，未被“触及”或“约束”。例如，用人工智能生成欺骗性内容干扰选举被认为是全球面临的重要挑战。二是大模型滥用对人类社会道德伦理秩序的漠视和扰乱。大模型充当着“人类大脑”或“世界模拟器”的角色，改变了知识的生产方式，缩短了知识生产的社会必要劳动时间，降低了人类获得所需文字、视频等生成内容的成本。大模型应用者从输入需求到输出结果，以及输出结果的未来用途等，全过程没有适用且有效的治理措施，这无疑是对人类社会道德伦理秩序的漠视和扰乱。

（四）大模型伦理失范所提出的治理挑战

大模型引起了技术海啸和秩序重构，对现有伦理治理体系提出了一系列挑战。现有的单要素和单环节的伦理治理模式、自上而下的伦理治理模式和以结果为导向的伦理治理模式已经无法有效防范和应对大模型伦理失范，体现为伦理治理的缺失或失效。

第一，单要素和单环节的伦理治理模式不能有效防范和应对大模型伦理失范。人机边界模糊与人机关系重构导致伦理治理主体范畴的扩张。人机智能交互和反馈式学习机制不断演进，模糊了人机边界，引起人机关系的重构，并产生“飞轮效应”，导致伦理治理范畴发生变化。一是大模型伦理治理对象发生变化。传统伦理治理范畴中的治理对象主要包括生成内容和服务提供者，大模型所引发的伦理风险则超越了以上治理对象，需将大模型研发者、训练者和应用者，训练数据的提供者、处理者和标注者等纳入伦理治理对象的范畴。二是大模型伦理治理主体尚未清晰界定，主体间责任则无法明确划分。以生成内容为例，大模型的生成内容是否存在伦理风险，取决于大模型在研发过程中是否存在伦理缺陷。因此，大模型研发者、服务提供者和应用者均在其中发挥了作用，对结果的输出负有责任，如果出现伦理失范，参与者均需要承担相应的责任，然而由于算法的不透明性和结果的不确定性，无法清晰界定不同主体的责任归属。综上所述，单要素和单环节的伦理治理模式呈现出一定程度的不适用性，无法有效防范和应对大模型伦理失范。

第二，自上而下的伦理治理模式不能及时防范和应对大模型技术迭代更新所引发的伦理风险。一是自上而下的伦理治理模式主要采用法律法规等方式治理伦理问题，与大模型技术迭代更新的速度相比具有一定的迟滞性和固定性，只能应对已经出现或可预见的伦理问题，却无法及时防范大模型技术迭代更新所产生的伦理风险。二是大模型具有技术不确定性所引发的生成内容及其社会影响的不确定性，自上而下的伦理治理模式具有一定的线性僵化特征，不能及时防范大模型对人类社会道德伦理秩序的冲击。三是以政府监管政策为主基调的自上而下的伦理治理模式，容易引起伦理准则抽象性与伦理实践之间的脱节，需要从伦理描述转向伦理应用[40]，探索新型伦理治理模式以避免大模型“野蛮生长”。

第三，以结果为导向的伦理治理模式不能有效防范和应对大模型生命周期的伦理问题。以《生成式人工智能服务管理暂行办法》为代表的监管政策体系，针对大模型及其相关应用的伦理治理注重对生成内容或信息的监管，责任主体的归属主要聚焦于生成式人工智能服务提供者。与其他人工智能技术相比，大模型技术的演进范式及其应用模式的转变引发的伦理失范现象不再局限于最终的生成内容，以结果为导向的伦理治理模式不能有效防范其伦理失范。一是大模型的技术创新范式发生变革，体现为开源式创新、过程式创新和不确定性创新。以结果为导向的伦理治理模式重心是大模型的生成内容或最终结果，不能有效应对大模型技术创新过程中存在的伦理风险。二是大模型的应用场景日趋多元化和开放化，导致伦理失范现象发生的潜在边界扩大，具有多点多发等特征。因此，以结果为导向的伦理治理模式已无法全面防范和应对大模型伦理失范。

四、大模型伦理治理框架的再构建

大模型在算法黑箱、数字鸿沟等伦理问题之上，衍生出了模型黑箱、数据版权侵害且主体责任难确定、冲击人类主体资格等伦理失范现象，对现行伦理治理体系提出了挑战，单要素和单环节、自上而下和以结果为导向的伦理治理模式已经呈现出治理上的缺位性、滞后性、固化性，以及政策与实践脱节等不足，亟须提出新的伦理治理框架以对大模型伦理失范进行防范和应对。在对大模型的属性、伦理失范缘起、关键要素解构和主要表现等进行分析，且对大模型伦理失范所提出的治理挑战进行阐述的基础上，本文以大模型生命周期为时间维度，以关键要素为核心，构建基于大模型关键要素的全生命周期伦理治理框架，如图2所示。

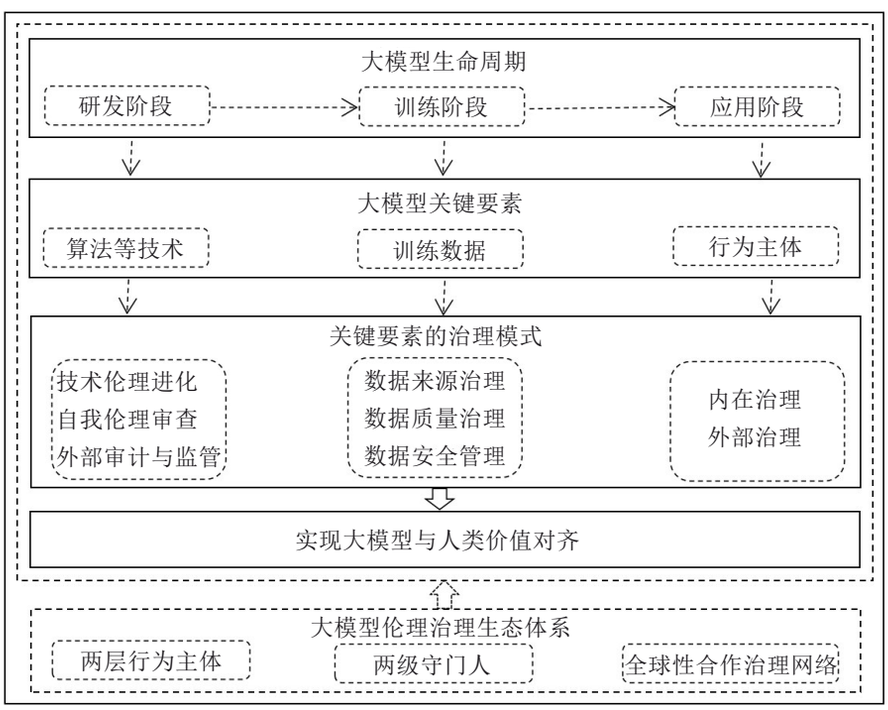

图2　基于大模型关键要素的全生命周期伦理治理框架

（一）基本思路

基于大模型关键要素的全生命周期伦理治理框架的基本思路是：一是覆盖大模型生命周期，强调伦理治理环节的完整性。大模型伦理治理应该覆盖其研发阶段、训练阶段和应用阶段，甚至延伸至退出阶段。二是重点关注大模型关键要素的伦理治理。算法、数据和人类这一行为主体是大模型的关键要素，也是伦理治理的重点。将大模型关键要素置于具体应用场景中，且强调伦理素养的内在提升和自我治理，有助于提高伦理治理的整体性、适用性和自下而上性。三是伦理治理目标是在人工道德代理（Artificial Moral Agent）的基础上，实现大模型与人类价值对齐或伦理对齐。与人类价值对齐，即通过人类的价值引导，大模型应与人类的价值观和社会道德伦理秩序相一致，《生成式人工智能服务管理暂行办法》中明确提出“坚持社会主义核心价值观”的具体要求。人类的价值观包括生存价值观、社会价值观和政治价值观。人类的社会道德伦理秩序主要指社会道德领域的正式制度和非正式制度。大模型与人类价值对齐体现为符合文化、符合社会和符合法规三个层次。符合文化体现为符合大模型所使用国家或地区的文化环境。符合社会体现为符合大模型所使用国家或地区的社会环境。例如，意识形态及政治倾向等。符合法规体现为对使用国家或地区的法律法规等制度的遵守与执行。前两项更多体现为对非正式制度的遵守，后一项

体现为对正式制度的遵守。四是重视大模型伦理治理的系统性与完整性。为配合基于大模型关键要素的全生命周期治理框架的有效实施，构建相应的伦理治理生态体系，也是该伦理治理框架的进一步细化与重要延伸。其中，大模型伦理治理生态体系中两层行为主体和两级守门人是对大模型关键要素的进一步细化，全球性合作治理网络是关键要素治理模式的延伸。例如，以大模型相关技术创新实现对大模型算法等技术的治理；发挥国际组织和各国政府的全局性及跨区域性伦理治理作用；社会舆论体系的伦理监督则是正式伦理治理制度的有效补充。

（二）大模型伦理治理的关键要素伦理治理

1.对技术的伦理治理

大模型是算法、模型、数据处理和标注、模型训练等技术的集合体。综合考虑大模型中算法、模型和参数之间的关联性和可分割性，将三者统一称为大模型技术。其中，算法伦理治理是重点。对大模型技术进行伦理治理，可以从三个维度展开：一是大模型技术的伦理进化。加强算法、模型、模型训练和调优、数据处理和标记等大模型技术创新。例如，以人类反馈强化学习（RLHF）、模型可解释性和OpenAI正在研发的模型审计等技术提高模型黑箱透明度、生成内容可解释性和伦理导向正确性等；在大模型中增加伦理数据训练模块、内置伦理自查或审计模型等伦理设计，减少伦理失范现象的出现和扩散。二是大模型技术的伦理自我审查、测试和评估。考虑到大模型技术具有商业秘密性和知识产权价值，且力求覆盖大模型技术的重要伦理风险敞口，可以采取模型卡制度，由大模型研发者创建模型卡，对大模型的参数、算法、训练数据、预期应用领域和用户群体等关键性信息进行登记和备案，以提高大模型的透明度和伦理规范性等。三是大模型技术伦理外部审计与监管。人工智能技术的备案、审查和安全评估等伦理治理措施已经推行，为加强大模型伦理外部审查的约束刚性，可对大模型技术采用全链条伦理审计制，针对不同技术可选择代码审计、非侵入式审计、众包审计、代理审计和抓取审计等方式[33]，以保证审计过程中的权责匹配性、机构独立性和知识产权保护等。

2.对数据的伦理治理

数据是大模型运行的驱动力，主要包括训练数据和用户数据两类。其中，大模型训练数据主要包括文本、图片、视频和（文本，视频）对等形式，在大模型预训练和调优阶段发挥作用。数据的来源、质量、合成技术和所含伦理内容等均会影响大模型所习得的伦理规则，并随着大模型应用场景的拓展体现出延伸性和扩散性。加强数据的伦理治理是大模型伦理治理的关键部分。

综合考虑大模型数据伦理风险敞口，可以从以下三个维度加强大模型数据的伦理治理：一是数据来源治理。对来源于公开数据、专门数据库等未经过相关处理的原始数据，应提高数据来源的可追溯性、正当性和规范性，避免出现数据版权侵害等伦理失范现象。此外，数据来源在一定程度上对数据质量和所包含的伦理理念等具有“隐性”保证作用，即来源正当的数据质量和伦理健康性更高。二是数据质量治理。大模型所使用数据的质量体现在是否存在误导性、歧视性等，对大模型、生成内容和应用者的道德伦理体系产生影响。加强对训练数据质量的检测与评估，并针对异常数据有相应的处理流程和办法，保证训练数据合乎伦理秩序和道德规范，避免低质量数据对生成内容和应用者的道德伦理体系产生负面影响。三是数据安全管理。数据安全和隐私保护是大模型运行的基石，以技术创新加强数据安全和隐私保护。数据存储和使用过程中均存在安全性要求，应加强隐私计算关键技术的突破。例如，联邦学习、多方安全计算、差分隐私和同态加密等，并加强关联领域的互联互通等，为数据安全和隐私保护提供一项新的技术选择，加上一把安全锁。从数据安全治理方式上，可引入专业型数据托管机构，将数据存储、管理和使用等职责分离，对数据处理和使用进行有效监督，提高数据安全性，也有利于政府对大模型训练数据进行监管。

3.对大模型行为主体的伦理治理

人类是影响大模型道德伦理体系的行为主体，是其伦理治理的关键要素。一是人类在人工智能大模型的研发、训练、部署和应用等环节均发挥主体作用，且贯穿始终。二是大模型的伦理风险触发和受影响主体是人类。三是大模型伦理治理的根本在于治人。按照大模型生命周期的不同阶段，将人类这一行为主体划分为大模型研发者、大模型训练者、数据提供者、数据处理者（主要包括数据预处理者、合成者和标记者）、大模型服务提供者和大模型应用者，针对以上行为主体进行伦理治理需综合考虑共性和差异性，即不同行为主体的伦理风险敞口存在差异。

其一，对大模型行为主体的内在伦理治理。其根本在于提升其伦理认知水平、伦理素质和专业技能。为从根源上实现伦理治理的目标，需要对大模型行为主体进行分层分类治理，切实提升伦理认知水平，且增强伦理风险防范能力。遵循伦理认知—伦理判断—伦理意图—伦理行为的伦理决策四阶段理论[43]，通过学校教育、公共宣传和在职培训等方式提高大模型行为主体的伦理认知层次和价值判断能力，正确引导伦理判断和伦理意图，降低伦理失范发生的概率，减轻伦理失范对人类社会道德伦理体系的消极影响。提升大模型研发者与训练者、训练数据提供者与处理者、大模型服务提供者等行为主体的伦理素质和专业技能，降低伦理失范发生的概率。大模型专业型行为主体是直接与大模型的算法、模型、参数和训练数据等接触的群体，拥有相对信息优势和空间优势，对大模型所习得的“道德伦理规则”有更深刻的影响，因而提升其伦理素质和专业技能更具迫切性且伦理治理成效更显著。专业型行为主体伦理素质的提升更多落脚于在职培训、职业道德准则学习等形式；专业技能的提升主要体现于模型黑箱的克服、伦理嵌入和伦理设计等领域，且将人类社会道德伦理规则贯穿于大模型研发、训练和服务提供整个过程，减轻大模型对人类社会道德伦理秩序体系的冲击。

其二，对大模型行为主体的外部伦理治理。推行许可证制、市场黑名单制和伦理失范问责制，以加强大模型行为主体的外部伦理治理。许可证制是在大模型推入市场之前，监管机构对大模型进行伦理评估或性能测试之后，允许符合标准的大模型进入应用市场；市场黑名单制根据大模型在应用中出现伦理失范的频率、影响范围和造成损失或伤害程度高等维度确定是否将大模型行为主体列入市场黑名单；伦理失范问责制主要从惩罚大模型相关行为主体、补救受影响主体损失并防范潜在伦理风险等层面对行为主体进行治理。

4.大模型应用场景、风险等级与伦理治理方式的匹配

欧盟《人工智能法案》提出，按照人工智能风险等级采取差异化治理方式。根据问题权变模型中道德强度的概念和维度，并分析了道德强度对伦理决策产生影响的路径。将风险等级、道德强度与场景理论融合，重点考虑大模型在不同应用场景中若发生伦理失范所造成社会影响的大小。大模型伦理风险等级的划分充分考虑道德强度的六个维度。其中，结果大小、社会舆论和效应可能性这三个维度的重要性高于其他维度，具体风险等级划分标准如表2所示。

表2　大模型伦理风险等级的划分标准

道德强度	结果大小	社会舆论	效应可能性	效应集中性	亲密性	时间即刻性
高	高	高	高	高	高	高
偏　高	高	高	高	低	低	低
中　等	高	高	低	低	低	低
偏　低	高	低	低	低	低	低
中　等	低	低	低	高	高	高
偏　低	低	低	低	低	高	高
偏　低	低	低	低	低	低	高
低	低	低	低	低	低	低

在借鉴欧盟分级分类治理模式、日本风险链模型、德国VCIO模型的基础上，将大模型主要应用场景划分为公共事业和商业服务两个类别，本文构建了基于应用场景的大模型伦理治理方式的匹配矩阵。①

五、大模型伦理治理生态体系的构建

为加强大模型伦理治理，实现大模型与人类价值对齐或伦理对齐的目标，在基于大模型关键要素的全生命周期伦理治理框架下，需要多元主体的共同参与，构建全方位的治理生态体系，主要包括推动大模型关键群体和企业两层行为主体的自我治理，发挥大型平台企业和政府两级守门人作用，形成以国际组织为核心、以技术治理为工具和以社会舆论为监督的全球性合作治理网络等子体系，加快大模型伦理治理“赤字”的绿化进程。大模型伦理治理生态体系包含大模型关键群体、大模型企业、数字守门人、国际组织和各国政府等多元主体的自治理、他治理和外部监管，能够实现对大模型道德伦理体系的影响因素相对全面的覆盖，不再局限于单要素和单环节，且关键群体、大模型企业等主体的自我治理和数字守门人的监管式守门人角色等是对以政府为主导的自上而下的伦理治理的有效补充。其中，重视发挥大模型研发者、大模型企业和数字守门人等主体的治理作用，实现从大模型研发、训练和服务提供等阶段均有相应的伦理自律或治理，既可以实现大模型生命周期的覆盖，又是伦理先行理念对以结果为导向的伦理治理模式的扩充和延伸，包括政府作为终极守门人在监管政策制定中应提高监管制度的全局性，力求将大模型生命周期统筹在内。

（一）推动两层行为主体的自我治理

大模型的行为主体可划分为关键群体和企业主体两个层级。其中，关键群体主要是大模型研发者，企业主体则是以企业形式存在的与大模型相关的经济组织。

第一，加强大模型关键群体伦理素养和专业技能的提升。人是产生伦理风险的根源，是大模型伦理治理的最终落脚点，大模型研发者是其中的关键群体，其拥有一定的信息优势和空间优势，伦理素养和专业技能对大模型原生道德伦理体系产生直接影响，是大模型伦理治理的第一线。大模型研发者的伦理自我治理可以从以下途径展开：一是提升自身伦理素养。以公共部门的公益性伦理教育或宣传为基础，以高等院校教育或继续教育的伦理课程为核心，以所从事具体领域的大模型伦理规范等制度文件的学习或培训为补充，强化大模型伦理教育和学习。二是提升自身专业技能。专业技能的提升具有长期积累性，且需要紧跟技术发展脚步进行迭代更新，大模型相关专业技能的提升可以从以下途径展开。首先，短期应以在职培训为重点，以实际工作需求为导向，注重理论知识与实践技能的结合，紧密结合技术发展前沿，提升大模型技术人员的专业技能。其次，中长期应以高等院校教育为重要途径，发挥个体自主性学习，以劳动力市场需求和个体职业发展为最终落脚点，培养专业型人才队伍。

第二，加强大模型企业的自我伦理治理。大模型企业主要包括大模型研发企业、大模型部署企业和服务提供企业，其自我伦理治理主要通过以下途径展开：一是强化伦理意识。伦理先行，将科技向善、负责任研究和技术创新等理念贯穿于企业文化与实践之中。发挥企业文化等非正式制度的引导作用，对企业伦理价值观具有正向影响。大模型的多模态、通用性特征和模型即服务（MaaS）的产业链模式等，要求加强企业间伦理倡议或公约等非正式伦理制度建设。二是完善自我治理的伦理制度体系建设，强化制度约束力且扩大覆盖面。大模型企业自我治理的伦理制度建设具有局部性特征，体现为集中于数据安全、隐私保护和用户个人信息安全领域，且以落实国家伦理监管政策和促进企业发展为主。大模型企业自我治理的伦理制度可从三个层面加以完善。首

① 基于应用场景的大模型伦理治理方式的匹配矩阵未在正文中列出，留存备索。

先，大模型技术规范、路线图和工具箱等技术伦理规范性制度建设。其次，大模型训练数据的相关伦理制度建设。最后，整体性大模型伦理风险防范和管理等制度建设。前两个层面强调制度的约束力和适用性，最后一个层面强调制度的约束力和覆盖面。正式制度是非正式制度的基础，对非正式制度具有补充作用，非正式制度是正式制度的延伸。三是变革组织架构。大模型企业加强自我伦理治理，且推动大模型伦理自律性制度有效施行，应进行相应的组织架构变革。首先，设置相对独立的大模型技术顾问委员会，其对企业大模型技术和应用是否符合伦理监管政策和自律性伦理制度进行监督，并就相关伦理问题提出建议和解决措施。其次，设置大模型伦理委员会，统筹不同部门以落实伦理治理战略，将企业已分散设置的数据隐私保护委员会、信息安全委员会、网络安全委员会等进行有机整合，保证大模型企业伦理治理的统一性和全局性；或者，保留现有伦理委员会，但需根据大模型关键要素、生命周期不同阶段和主要伦理失范现象等调整伦理审查、监督等方向和内容。最后，在业务部门设置大模型伦理专员或工作小组等，负责企业大模型伦理战略、自律性制度和非正式制度在大模型技术研发、产品设计和生产过程的具体执行和引导等。四是创新自我伦理治理实践，主要包括加强对外披露自我伦理治理相关内容，以及加强大模型伦理自治技术或工具的应用及创新。目前，大模型企业主要以企业社会责任报告、ESG报告和道德准则等专项报告为载体，需进一步将伦理自治相关披露内容从负责任算法、负责任人工智能延伸至负责任大模型。大模型伦理审查、测试、评估和补救等技术或工具有助于降低伦理失范现象的出现，提高大模型的安全性、可靠性和稳健性等。

（二）发挥两级守门人作用

第一，发挥数字守门人的伦理治理作用，不断迈向平台中立。欧盟《数字市场法》对数字守门人进行了详细界定，其主要指超级平台企业和大型平台企业。数字守门人可以凭借其数据、算法和市场等优势或权力获取一定的伦理治理影响力，且以强大的网络效应形成以平台为中心的联合企业体系，对这一体系的伦理治理发挥一定作用。因此，数字守门人兼具自律式守门人[44]和监管式守门人[45]的双重角色，介于政府与其他企业主体之间。从平台企业加重义务和社会责任履行的角度出发，进一步发挥数字守门人作用可从以下途径展开：一是发挥超级平台企业和大型平台企业的自律式守门人作用。数字守门人在履行政府伦理监管政策并承担相应违规责任的基础上与用户之间建立良好的信任关系且延伸至社会信任体系。首先，形成用户数据的权属清晰、数据安全和隐私保护等基础性信任关系。在现行用户隐私设置或信息收集等基础上，明确用户对数据的所有权、携带权和被遗忘权等，在用户数据的存储、使用和销毁等过程中保护数据安全和用户隐私等，尤其是杜绝用户数据滥用和非法交易等。其次，在默认设置、访问渠道和用户推荐等平台应用中形成透明、非歧视性和破除“信息茧房”等过程性信任关系。例如，部分用户隐私声明存在一定的形式性，将数据使用权和实际控制权一并让渡的同时，也将隐私侵犯、数据滥用等伦理风险敞口暴露，过程性信任关系不可或缺。最后，在生成内容等服务提供或输出结果中形成结果性信任关系。二是发挥超级平台企业和大型平台企业的监管式守门人作用。发挥平台企业对用户的部署领域、输入内容和生成内容等方面的伦理“准监管”作用，凭借其控制力对用户违背社会伦理规则、监管政策和平台自制监管规则的行为或结果等采取干预措施或自行裁罚等，在联合企业体系中产生威慑力和警示性。

第二，发挥政府终极守门人作用，加快伦理监管制度供给和监管机构设置步伐，创新监管工具和方式。政府是大模型伦理监管政策的主要供给主体，需逐步完善伦理监管制度体系。为应对大模型伦理失范对现行伦理监管制度提出的挑战，应提高制度的前瞻性、全局性、灵活性和适用性。一是提高伦理监管制度的前瞻性。紧跟大模型技术发展前沿，政府所属科研机构应加强对大模型伦理的相关研究，发挥其建言献策的作用，力求以潜在伦理风险为导向，提前布局伦理监管制度建设。二是提高伦理监管制度的全局性，力求覆盖大模型关键要素及生命周期。以大模型为

监管对象，以关键要素为监管核心，以生命周期为监管跨度，改变数据、算法和服务提供者三者相分离的监管局面，形成系统性的伦理监管制度体系。三是提高伦理监管制度的灵活性和适用性。以大模型应用场景、风险等级与伦理治理方式相匹配为原则，在“分领域监管”理念的基础上，引入“分场景监管”理念，实行重点场景、高风险等级重点监管，实现伦理监管制度制定的灵活性和适用性。其中，将法律法规等形式的硬性监管制度布局于重点场景、高风险领域，明确大模型关键要素和行为主体的伦理责任，划出清晰的法律红线和底线，提高制度约束的刚性。相反，对于非重点场景、低风险领域，则主要布局意见和办法等形式的软性伦理监管制度。

另外，政府应设置相应的大模型伦理监管机构，并注重不同层级监管机构之间合作与交流机制的形成和优化。一是设置双层的伦理监管机构。考虑到单一制国家、复合制国家和联盟等国家或联合体结构的差异，在中央（联邦）政府、联盟总部层面设置一个综合性大模型伦理监管机构，以执行伦理审计、审查、评估、调查、救济和监测伦理监管制度的实施等相关监管职能为主。其中，伦理审计是在现行“备案制”基础上的延伸，是伦理监管约束力增强的体现。在地方政府（州）、成员国层面设置相应的伦理监管机构，与上述伦理监管机构之间存在一定行政隶属关系，职能配置上相似。二是形成或优化双层伦理监管机构的运行机制，核心为统一协调与合作交流，以保障伦理监管制度的执行、防范伦理失范的发生和相关伦理失范事件的处理等。此外，探索在监管机构内部设置一名首席大模型伦理治理官，对该机构的伦理监管事宜进行统筹和规划，形成上下统一且专一的伦理监管机构体系。优化大模型伦理监管队伍，创新监管工具。大模型等生成式人工智能应用属于科技范畴，伦理监管属于社会科学范畴，大模型伦理监管属于两者交叉领域，单纯依靠其中一个学科领域的人员难以实现兼顾全面性和适用性的伦理监管。因此，在大模型伦理监管人才队伍中应吸纳专业技术人员，包含实际操作技术人员和行业技术专家。此外，在监管沙盒、工具包“A. I. Verify”等基础上，政府仍需加快大模型伦理监管工具的创新。

（三）形成全球性合作治理网络

大模型伦理治理是全球性治理难题，需加强全球及区域间伦理治理合作与交流，以全球性或区域性国际组织为引领，以政府间合作与交流为重要形式，以制度建设和治理技术创新为具体表现，同时，社会舆论体系应把好伦理监督关，共同形成且织密全球性合作治理网络。形成全球性合作治理网络，推进大模型伦理治理“赤字”的绿化进程。一是全球性国际组织对伦理治理网络进行整体性布局，多边或双边等区域性国际组织进行符合当地需求的布局，以倡议书、规范和标准等不具有强制性的制度为主。除制度建设外，充分发挥国际组织的号召力和组织协调能力，以大模型技术发展和伦理治理为主题开展研究和交流，为制定伦理治理制度等提供智力支持，增强不同国家和区域的合作与交流，且逐渐形成伦理治理合力等目标。二是加强政府间及区域间伦理治理合作与交流。以训练数据、用户数据和算法等大模型关键要素跨境流动、国际数字贸易等现实需求为抓手，以贸易合作、技术合作合同等为切入点，以增强数据安全、技术保密性等为短期目标，将大模型伦理治理相关内容内嵌其中，不断推动伦理治理从双边互认互信、多边互认互信至全球性互认互信。三是创新伦理治理技术，实现以技术治理技术。增强全球性和多边性技术组织的创新能力，大模型底层支撑是人工智能技术，模型黑箱、数据版权侵害等伦理失范现象的根源在于现行技术存在人类未解之谜或未涉足领域。首先，加强大模型技术自身的创新。联合全球性、区域性或各国技术组织，以关键性算法、数据处理技术和隐私计算等为主要攻关领域，克服大模型生成内容不确定性对社会伦理体系的冲击。其次，在大模型技术中添加伦理训练和优化部分，使大模型原生道德伦理体系和衍生道德伦理体系均与人类价值对齐。四是发挥社会舆论体系的伦理监督作用。社会公众是大模型应用者，是受益者也是受损害者，是大模型伦理监督最广大的群体，是在伦理治理制度体系、治理机构和治理技术等措施之外的有效补充，具有柔性化、灵活性、及时性和覆盖面广等特征。

六、总结与展望

大模型的问世将人工智能技术从专用性、单一模态演进至通用性、多模态，助力人类生产生活数字化和智能化转型提速，与此同时引发伦理失范及治理难题。本文在对现有研究进行梳理和分析基础上，从生产力、生产工具相关理论对大模型属性及伦理失范缘起进行探究，发现作为生产力的大模型存在技术非中性，作为生产工具的大模型内嵌人类伦理，生产力和生产工具双重属性的大模型将人类伦理与机器伦理糅合其中，大模型区别于传统生产力和生产工具，其自身和生成内容均内嵌伦理。继而从技术和生命周期角度对大模型伦理失范的关键要素进行解构，发现大模型伦理失范的技术起点是算法，关键中介是数据，行为主体是人类。针对大模型诸如模型黑箱、数据版权侵害且主体责任难确定、冲击人类主体资格等伦理失范现象，对单要素和单环节、自上而下和以结果为导向的伦理治理模式提出挑战，以大模型生命周期为时间维度，以关键要素为核心，本文构建了基于大模型关键要素的全生命周期伦理治理框架，以有效防范与治理大模型伦理失范。为推动大模型伦理治理新框架的有效运行，本文构建了包含两层行为主体的自我治理、两级守门人和全球性合作治理网络等子体系的伦理治理生态体系，加快大模型伦理治理“赤字”的绿化进程。

综合考虑大模型的技术演进、伦理失范特征、伦理治理进程等相关研究现状，未来可在以下领域进行深入研究：一是以大模型应用场景为研究切入点，细化大模型应用场景分类，针对不同应用场景采取不同的伦理治理模式，实现大模型精细化治理、精准化治理。二是从大模型技术创新、迭代更新的视角，以算法黑箱、模型黑箱等关键伦理问题为核心，从技术和理论两个层面开展如何实现以算法治理算法、以模型治理模型等相关研究。三是从法律层面，针对大模型伦理失范中不同行为主体的责任归属、责任界定和处罚措施等，开展系统性研究。

参考文献：

[1] 张辉，刘鹏，姜钧译，等.ChatGPT：从技术创新到范式革命[J].科学学研究，2023，41(12)：2113-2121.

[2] FOFFANO F，SCANTAMBURLO T，CORTES A.Investing in AI for social good：an analysis of European national strategies[J]. AI & society：the journal of human-centered systems and machine intelligence，2023，38(2)：479-500.

[3] ROBERTE H，COWLS J，MORLEY J，et al.The Chinese approach to artificial intelligence：an analysis of policy，ethics，and regulation[J]. AI & society：the journal of human-centered systems and machine intelligence，2021，36(1)：59-77.

[4] 赵志耘，徐峰，高芳.关于人工智能伦理风险的若干认识[J].中国软科学，2021(6)：1-12.

[5] 孙蒙鸽，韩涛，王燕鹏，等.GPT技术变革对基础科学研究的影响分析[J].中国科学院院刊，2023(8)：1212-1224.

[6] 张凌寒.生成式人工智能的法律定位与分层治理[J].当代法学，2023(7)：126-141.

[7] 王钰，唐要家.人工智能应用如何影响企业创新宽度?[J].财经问题研究，2024(2)：38-50.

[8] FAWAZ Q.ChatGPT in scientific and academic research：future fears and reassurances[J].Library hi tech news，2023，40(3)：30-32.

[9] 陈兵.促进生成式人工智能规范发展的法治考量及实践架构——兼评《生成式人工智能服务管理暂行办法》相关条款[J].中国应用法学，2023(4)：108-125.

[10] WEI J，WANG X，SCHUURMANS D，et al.Chain-of-thought prompting elicits reasoning in large language models[EB/OL].(2022-01-28)[2024-02-26].https://arxiv.org/abs/2201.11903.

[11] MCLENNAN S，FISKE A，TIGARD D，et al. Embedded ethics：a proposal for integrating ethics into the development of medical AI[J]. BMC medical ethics，2022，23(1)：1-10.

[12] JULES W，QUCHEN F，SAM H，et al.A prompt pattern catalog to enhance prompt engineering with ChatGPT[EB/OL].(2023-02-21)[2024-02-24].https://arxiv.org/pdf/2302.11382.pdf.

[13] 滕妍，王国豫，王迎春.通用模型的伦理与治理：挑战及对策[J].中国科学院院刊，2022，37(9)：1290-1299.
[14] 邓悦，许弘楷，王诗菲.人工智能风险治理：模式、工具与策略[J].改革，2024(1)：144-158.
[15] 张欣.面向产业链的治理：人工智能生成内容的技术机理与治理逻辑[J].行政法学研究，2023(6)：43-60.
[16] 王沛然.从控制走向训导：通用人工智能的"直觉"与治理路径[J].东方法学，2023(6)：188-198.
[17] 徐伟.论生成式人工智能服务提供者的法律地位及其责任——以ChatGPT为例[J].法律科学(西北政法大学学报)，2023，41(4)：69-80.
[18] 李石勇，卜传丞.科技伦理治理中法治不能缺位[N].中国社会科学报，2022-06-14(2).
[19] 赵精武，王鑫，李大伟，等.ChatGPT：挑战、发展与治理[J].北京航空航天大学学报(社会科学版)，2023(3)：188-192.
[20] LEANDRO M.Editorial：ChatGPT and the ethical aspects of artificial intelligence[J]. Revista de gestão，2023，30(2)：110-112.
[21] OZKAYA I.Application of large language models to software engineering tasks：opportunitier，risks，and implications[J]. IEEE software，2023，40(3)：4-8.
[22] SOMEH I A，DAVERN M，BREIDBACH C F，et al.Ethical issues in big data analytics：a stake-holder perspective[J]. Communications of the association for information systems，2019，44(34)：718-747.
[23] 续继，王于鹤.数据治理体系的框架构建与全球市场展望——基于"数据二十条"的数据治理路径探索[J].经济学家，2024(1)：25-35.
[24] 张欣.生成式人工智能的数据风险与治理路径[J].法律科学(西北政法大学学报)，2023，41(5)：42-54.
[25] 丁晓东.论算法的法律规制[J].中国社会科学，2020(12)：138-159+203.
[26] 刘艳红.生成式人工智能的三大安全风险及法律规制——以ChatGPT为例[J].东方法学，2023(4)：29-43.
[27] 肖红军.算法责任：理论证成、全景画像与治理范式[J].管理世界，2022，38(4)：200-226.
[28] 苏宇.算法规制的谱系[J].中国法学，2020(3)：165-184.
[29] 张凌寒.深度合成治理的逻辑更新与体系迭代——ChatGPT等生成型人工智能治理的中国路径[J].法律科学(西北政法大学学报)，2023，41(3)：38-51.
[30] 周文，许凌云.论新质生产力：内涵特征与重要着力点[J].改革，2023(10)：1-13.
[31] 周文，何雨晴.新质生产力：中国式现代化的新动能与新路径[J].财经问题研究，2024(4)：3-15.
[32] ROBERT，K.Algorithm= Logic + Control[J]. Communications of the ACM，1979，22(7)：424-436.
[33] 陈雄燊.人工智能伦理风险及其治理——基于算法审计制度的路径[J].自然辩证法通讯，2023(10)：138-141.
[34] 孙伟平.人工智能与人的"新异化"[J].中国社会科学，2020(12)：119-137.
[35] 杨双彪.人工智能的伦理困境及其规避进路——基于马克思主义科技观分析[J].理论观察，2023(6)：52-56.
[36] EDEN R，BURTON-JONES A，CASEY V，et al.Digital transformation requires workforce transformation[J]. MIS quarterly executive，2019，18(1)：1-17.
[37] KIM E S.Deep learning and principal-agent problems of algorithmic governance：the new materialism perspective[J]. Technology in society，2020，63(11)：101378.
[38] CRNKOVIC G D，ÇÜRÜKLÜ B.Robots：ethical by design[J]. Ethics and information technology，2012，14(1)：61-71.
[39] MILLER K W.Moral responsibility for computing artifacts：the rules[J]. IT professional，2011，13(3)：57-59.
[40] 孙保学.人工智能算法伦理及其风险[J].哲学动态，2019(10)：93-99.
[41] JUNAID Q.Toward accountable humancentered AI：rationale and promising directions[J]. Communication and ethics in society，2022，22(2)：329-342.
[42] PEARL J.The seven tools of causal inference，with reflections on machine learning[J]. Communications of the ACM，2019，62(3)：54-60.
[43] REST J R.Moral development，advances in research and theory[M]. New York：Prager，1986：23.
[44] 候利阳.论互联网平台的法律主体地位[J].中外法学，2022，34(2)：346-365.
[45] 解正山.约束数字守门人：超大型数字平台加重义务研究[J].比较法研究，2023(4)：166-184.

Theoretical Deconstruction and Governance Innovation of Ethical Misconduct in Large Model

XIAO Hong-jun[1], ZHANG Li-li[2]

(1. Institute of Industrial Economics, Chinese Academy of Social Sciences, Beijing 100006, China;
2. Institute of Digital Economy, Beijing Academy of Science and Technology, Beijing 100089, China)

Summary: Large model is at the forefront of artificial intelligence (AI) technology, becoming a "universal simulator" connecting the physical world and the digital world. Large model has achieved breakthrough and disruptive technological innovation, bringing about changes in production methods, technological innovation paradigms, content generation methods, and human-machine relationships. At the same time, it has also triggered a series of ethical misconduct phenomena, becoming a global governance challenge. Existing literature on how to prevent and govern ethical misconduct in a large model still needs to be further studied and enriched, and there are shortcomings such as weak targeting and applicability. In addition, the existing ethical governance model is no longer able to effectively address the issue of ethical misconduct in a large model, we urgently need to propose a new ethical governance framework.

This article provides a theoretical deconstruction of ethical misconduct in a large model and finds that the origin of the ethical misconduct of a large model lies in its distinguishing characteristics from traditional productive forces and production tools, such as technological non-neutrality, embedded human ethics, and the mixing of machine ethics and human ethics. The starting point of the ethical misconduct of a large model is the algorithm, the key intermediary is data, and the behavioral subject is human. The ethical misconduct phenomena caused by a large model, such as black-box models, data copyright infringement, and the impact on human subject qualifications, which poses a challenge to the current ethical governance system. To address the challenge, with the life-cycle of a large model as the time dimension and key elements as the core, this article constructs a full life-cycle ethical governance framework for key elements of a large model. At the same time, this article aims to promote the effective implementation of the new ethical governance framework, and build an ethical governance ecosystem that includes subsystems such as two levels of self-governance, two levels of digital gatekeepers, and a global collaborative governance network.

This article expands previous literature from the following aspects. Firstly, it deconstructs the attributes and key elements of a large model in theory, which is a re-understanding of the essence of generative AI technologies such as a large model. Secondly, the theoretical deconstruction of the causes of ethical misconduct in a large model is conducted to further reveal the role of key elements such as algorithms, data, and humans in ethical misconduct. Thirdly, the phenomenon of ethical misconduct in a large model is summarized and extended based on the phenomenon or risk of ethical misconduct in AI, such as "model black boxes". Fourthly, this article proposes a new governance framework for ethical misconduct in a large model. Beyond the current single factor, result-oriented governance model, this article constructs an ethical governance framework with key elements as the core, combined with the life-cycle of a large model, and constructing a corresponding ethical governance ecosystem to effectively prevent and govern ethical misconduct in a large model.

Key words: large model; theoretical deconstruction; ethical governance

（责任编辑：尚培培）

[DOI]10.19654/j.cnki.cjwtyj.2024.05.002

[引用格式]肖红军，张丽丽．大模型伦理失范的理论解构与治理创新[J]．财经问题研究，2024(5)：15-32.