■ 人工智能与社会发展研究

论生成式人工智能的技术创新伦理周期

——以 ChatGPT 为例

雷宏振，刘　超，兰娟丽

(陕西师范大学，国际商学院 陕西 西安 710119)

摘　要：生成式人工智能在推动技术创新突破性应用的同时，也带来了许多社会伦理风险：隐私暴露与信息安全控制风险；虚假信息派生所带来的信息传播风险；算法偏见所带来的决策和政策风险以及责任界定模糊风险。运用“索洛悖论”理论，将伦理进步看作是科技创新投入的增函数，可以发现生成式人工智能对伦理推动的“技术创新索洛周期”。基于此，要跨越生成式人工智能“伦理风险鸿沟”可以采取以下对策：健全生成式人工智能法律规制，推动构建生成式人工智能治理体系；加强平台监管，完善流程标准化管理，规范保障生成式人工智能安全发展；落实优化生成式人工智能可持续发展环境，探索更多行业重塑的可能。

关键词：生成式人工智能；技术创新；索洛悖论；伦理风险

中图分类号：TP18；F499　　文献标识码：A　　文章编号：1672－4283(2024)01－0097－11

收稿日期：2023－09－25　　DOI：10.15983/j.cnki.sxss.2024.0101

作者简介：雷宏振，男，陕西合阳人，经济学博士，陕西师范大学国际商学院教授，博士研究生导师。

一、引　言

ChatGPT 的出现标志着生成式人工智能取得突破性进展，对人类社会和经济发展产生了广泛而深刻的影响[1-2]。但随之而来出现了一些新型犯罪方式，如盗用私人信息和生成人脸诈骗、利用系统漏洞非法获取相关服务器所属企业的支付密钥等伦理问题迅速引发全球范围内的关注和担忧。2023 年 3 月 30 日，联合国教科文组织(UNESCO)总干事奥德蕾·阿祖莱发表声明，呼吁各国尽快实施该组织于 2021 年 11 月通过的《人工智能伦理问题建议书》，旨在为生成式人工智能发展设立伦理标准以最大限度发挥人工智能优势，降低其带来的伦理风险①。

如何看待生成式人工智能在推动社会经济发展过程中所引发的社会伦理道德冲突？针对这一问题学者们已开始关注[3-5]，但缺少经济学理论层面的解释。事实上，基于现象的研究与溯因逻辑在问

① 见 https://www.chinadailyasia.com/article/323365。

题解决方面尤其有效[6-8]。因此,从生成式人工智能应用引发伦理困境这一现象出发,追溯生成式人工智能与社会伦理的发展规律,对提出和实施符合人类价值观的解决方案、规范引导“生成式人工智能”科技更健康、有序地服务于社会经济发展具有重要的科学意义。

根据索洛悖论,技术创新从出现到效用的显现存在技术扩散周期,即在新技术创新初期产出不会出现增长效应,相反还可能出现一定程度上的下滑。这是由于新技术的采纳和应用效果具有一定的时滞性,技术创新效应具有“U 型”特征。生成式人工智能是人工智能技术领域的 1 项颠覆性创新,具有里程碑的意义,因技术创新的“索洛悖论”特征,我们可以推断其推动的伦理进步也应该遵循类似的发展规律,即存在技术创新伦理周期。这一过渡期既是技术的升级更新、监管和约束机制及时补位和调整的过程,也是公众科学素养提高的过程。但在技术应用初期,部分人会利用技术创新周期与滞后的监管体系这一漏洞实施不道德行为,引发伦理困境。基于此,本文以 ChatGPT(Chat Generative Pre-trained Transformer,生成式预训练语言转换器)应用为例,首先梳理了其技术创新效应以及公众对可能出现的伦理问题的担忧,在此基础上讨论以下几个问题:生成式人工智能是否存在技术创新伦理周期?若是,本文的目的是:以具有代表性的生成式人工智能产品 ChatGPT 为研究对象,提出其遵循“索洛悖论”的发展规律,进一步将伦理进步看作是生成式人工智能科技创新投入的增函数,提出 ChatGPT 对伦理的促进作用同样具有“U 型”特征,当前应用过程暴露的伦理问题是“技术创新伦理周期”的阶段性特征,是技术创新周期与系统性规范体系动态演进过程的正常现象,为公众辩证地看待生成式人工智能发展提供参考,并就如何推动此类科技进步走向健康发展的轨道提出建议。

二、ChatGPT 技术创新的伦理风险

(一) ChatGPT 的技术创新效应

ChatGPT 是人工智能技术驱动的聊天机器人。基于大型语言模型并利用全球公开数据集开展深度学习训练,ChatGPT 能够采集多种情境下的语言模式,在几秒内理解用户输入的对话式指令并给出交流式的高质量反馈。与已有的人工智能技术相比,ChatGPT 不仅可以“理解”人类自然语言,“记住”训练期间获得的大量事实,还可以基于“记住”的知识生成高质量的内容,几乎在任何需要处理和理解自然语言的领域都可以发挥关键作用,具有进一步释放经济增长活力的潜力[9]。

其一,ChatGPT 凭借其强大的自然语言处理能力提高了生产效率。ChatGPT 以直观、友好的方式处理和响应用户问题,并根据用户查询的上下文内容生成类似人类的响应,自动化完成相对常规、耗时的翻译、起草电子邮件、撰写简历、编写计算机代码、检查程序等任务。在这种情况下,工作对人的依赖性降低,可以更少的人完成相同的工作,服务将变得更便宜,消费需求可能会扩大,有助于推动就业和提高生产力。

其二,ChatGPT 利用大型语言模型和非结构化数据优化现有模型,大大提高了反应时效性和模型预测靶向性。ChatGPT 的本质是基于深度学习框架的大型神经网络模型,具备预训练语言模型(Pre-Trained Language Model)和上下文学习模型(In-Context Learning),可用于全球范围内的数据、知识和技术资源的输入模型训练,其中上下文学习模型嵌入了思维链提示和指令微调的新技术。通过对现有模型和算法进行质量校正和模型优化,可以加速知识迭代和提升预测准确性,减少效率损失[10-11]。在教育行业,ChatGPT 能够将复杂的科学理论和概念提炼成简单的语言,增强学生的理解记忆能力。在生成完整解决方案的基础上,将复杂问题分解并构建任务解决流程。ChatGPT 还可以评估学生作业或论文中的表现并提供反馈和建议,既可以提高教师评估学生和学生自我评估的效率,也有利于跟踪学习、消除知识短板[12]。

其三,ChatGPT 与其他要素投入互补,赋能产业发展和衍生新兴关联产业。从技术发展来看,ChatGPT 相关的隐私计算、实时计算、硬件变革等技术发展推动了大数据技术和产品的升级迭代。其利用大型神经网络和生成式预训练转换器(Generative Pre-trained Transformer,GTP) 从全球范围内的数据、知识和技术资源(包括文章、社交媒体海报、论文、计算机编程代码和电子邮件形式的非结构数据) 中提取信息,"喂养"数据来源广泛,能够满足用户多元化需求和兴趣。相应地,ChatGPT 既可以引入数据、算法等新要素形成部分新的高效率产业,推动经济快速增长,又可以与现有技术、产业形成互补,实现跨界融合地应用,提高技术效率[10] [13]。如夏克德和惠特尼通过一项在线实验研究了 ChatGPT 的生产力增长效应,将特定职业的激励写作任务(撰写新闻稿、简短报告、分析计划和精致的电子邮件等) 分配给 453 名受过大学教育的专业人员,并随机安排一半的被试使用 ChatGPT 作为辅助工具。结果显示,与控制组相比,使用 ChatGPT 的被试平均花费时间约减少了 40%,任务完成质量提高了 18%。ChatGPT 结合专业知识明显有助于激发行业发展的新动力。[14]

(二) ChatGPT 的伦理风险

ChatGPT 的应用引起人类社会的深层次变革,为未来经济社会发展带来无限的机遇与可能。但与此同时,其大规模式自传播很快引发了各界关注与担忧,核心焦点之一就是可能引发的伦理问题,主要体现在以下几个方面:

1. 隐私暴露与信息安全控制风险

ChatGPT 收集和处理数据时,未征得用户授权,或者超范围使用,存在个人信息和隐私泄露的风险,使得一些违法分子有空子可钻。其一,用户隐私信息暴露风险。ChatGPT 的监督策略模型训练数据来源包括用户的输入数据。在使用 ChatGPT 进行聊天时,用户可能会无意识地提供额外的私人信息,这都可能被收集自循环用于进一步训练模型。用户享受其便利的同时,也面临隐私暴露风险。其二,版权纠纷风险。ChatGPT 中大多数生成模型都是在未经许可或无偿的情况下从互联网上收集材料进行训练的,数据来源广泛,包括诗歌、法律文件、自然对话、博客和电子邮件等,这些数据不可避免地包含了受版权保护的信息,因此,ChatGPT 生成的回答存在过多地引用他人工作,或者未经许可保留训练数据集中于特定作品的知识隐患。其三,网络犯罪安全漏洞。ChatGPT 应用场景广泛、操作简单、可定制性强,基本上可以让零编码技能的人成为网络罪犯,大幅度降低原本需要分工进行的链条式犯罪的难度。一方面,ChatGPT 可以帮助黑客和网络犯罪分子发现网站上的安全漏洞,更快更容易地生成网络攻击脚本。另一方面,ChatGPT 允许零编码技能的人开发恶意软件,从编写令人信服的网络钓鱼电子邮件到编写恶意代码,再到规避常见的网络安全检查,这为潜在的诈骗者、跟踪者、间谍、犯罪分子和恐怖分子提供了可乘之机。

2. 虚假信息派生所带来的信息传播风险

ChatGPT 利用"反馈式强化学习"技术吸收用户的反馈进行机器训练,能够满足用户主体的多元化请求。但是 ChatGPT 有时也会给出似是而非但不正确的答案[15],为错误信息的传播和泛滥提供了可能[5]。ChatGPT 生成的答案存在事实或逻辑错误的风险。事实错误可能是由于训练数据本身存在错误或噪声[16]。首先,ChatGPT 的训练数据来源多元化,包括但不限于诗歌、自然对话、博客和电子邮件等,这些数据并不要求逻辑严谨或符合科学规律,数据源头本身可能存在错误。其次,ChatGPT 实现向人类看齐的能力依赖于手动标记和分类数据的质量,目前它还没有能力在人工标记之外自动区分和确认模糊的数据。如果出现人工操作失误导致不恰当的内容过滤不当,那么它很有可能会输出错误信息,这给 ChatGPT 在真实场景中的应用带来了不可信赖的障碍[15]。再次,数据驱动的深度学习模型的内部逻辑和机制仍然是"黑匣子",关于为什么会产生当前答案的猜测既不能被证明,也不能

被证伪。最后,ChatGPT 训练集的知识仅限于 2021 年 9 月之前的信息,这意味着任何在该日期之后发生的事件或新知识都不会被模型完全理解和处理,这可能导致部分结果存在偏差[17]。ChatGPT 产生的错误信息,会误导网络弱势群体(如老人、妇女、儿童、数字素养有限的个人等)和虚假信息的广泛传播。例如从浙江绍兴警方摧毁的一个利用 ChatGPT 制作虚假视频获取流量的团伙案件可以看出,该团伙利用 ChatGPT 自动生成"脚本"和相关视频素材后一键生成有音乐、字幕的"上虞工业园区发生火灾"虚假视频,短时间内视频浏览量迅速攀升,给当地造成不小的影响①。尽管目前提出的技术(工具形成器和插件②),可以部分缓解事实错误的问题,并且已经展开了大量的工作,但在各项作业和领域中的有效性仍需不断的观察和探索。

3. 算法偏见所带来的决策和政策风险

开发大语言模型的一些信息存在性别不平等、种族歧视、文化偏见等。开放人工智能(OpenAI)公司在决定公开发布 ChatGPT 时,试图通过两种方式避免和过滤敏感类话题的传播:一是将数据库限制在 2021 年以前;二是采取筛选模型训练模式和训练文本。但实际上,过滤功能的实现需要特定的员工给敏感文本贴上标签,因而 ChatGPT 的训练数据中仍嵌入如特定文化的优越性等偏见或想法,存在加剧性别不平等、种族歧视和文化偏见的风险[18]。

4. 责任界定模糊风险

基于深度学习框架的大型神经网络模型,ChatGPT 的工作原理和运行机制是不透明的,将引发不可解释问题以及责任主体认定模糊风险。"黑箱效应"使得用户很难理解系统是如何做出决定和产生答案的,也不清楚是谁要对生成数据的准确性负责。这引起各界关于 ChatGPT 的法律人格与生成内容认定的探讨[19]。例如,在学术创作中,部分学者认为可以根据其重大贡献将 ChatGPT 类人工智能列为共同作者并明确其工作内容[20]。《自然》杂志则提出相反的观点,原因在于,任何作者的归属都带着工作的责任,而人工智能不能承担这种责任,因此没有任何人工智能工具将被接受为研究论文的作者。但是如果研究人员在论文写作中使用了人工智能工具,应在适当的部分记录使用情况,保证研究方法的透明度。还有部分观点认为科研人员不应使用 ChatGPT 撰写科学研究的任何部分,但是可以在人类监督下使用,以确保科学工作的完整性和原创性[21]。

三、ChatGPT 的技术创新伦理周期:基于"索洛悖论"视角的解释

(一)"索洛悖论"理论及其应用

1. "索洛悖论"理论

"索洛悖论"指信息技术的大量应用却未带来劳动生产率或全要素生产率在统计数据上的增长[22],其实质是 1 项技术创新从应用到效用显现存在过渡期。自"索洛悖论"提出以来,学者对其存在性[23-24]、发展脉络、产生原因[25],在不同层面的表现及跨越路径[26]进行了探索性的总结与梳理。就信息技术革命呈现的"索洛悖论"发展规律,多数观点支持"索洛周期"时滞效应假说,即信息技术与以往的技术革命一样具备提高生产率的潜能,但是这种潜力的显现需要一段时间[27]。从信息技术的生命周期角度理解:信息技术与经济增长呈"U 型"关系。整体上,信息技术对于效能的提高具有促进作用,但随着技术生命周期的演进,关系的强度也会随之变化,表现为:在技术引入阶段高度不稳

① 见《绍兴警方侦破利用 ChatGPT 技术团伙制作虚假视频案》,中国新闻网,2023 - 07 - 05,https://www.chinanews.com.cn/sh/2023/07-05/10037095.shtml。

② 见 https://openai.com/blog/chatgpt-plugins。

定,增长阶段为强正,成熟阶段为弱正,下降阶段为负。还有部分学者将技术革新与经济增长效应的“时间差”归因于统计学上的测量误差[28-29]和管理不当导致的资源效率低下[29]。随着新一代信息技术时代的到来,有学者考察了人工智能和大数据等对生产率的贡献。陈楠和蔡跃洲回顾了中国省域面板数据,发现人工智能技术仅对经济增长规模产生了促进作用,对增长速度和效率提升并不显著,呈现新“索洛悖论”特征。[30]

2. “索洛悖论”的延伸解释力

有学者通过改变解释变量或被解释变量,将“索洛悖论”应用于各个行业领域,对其内涵加以延伸和补充。在不同时间、不同地域或不同行业层面,“索洛悖论”被证实广泛存在,但存在差异。大部分制造业企业的研究都支持了悖论的存在[31]。例如方颖和余兴锦基于“索洛悖论”的研究视角,以数字化投入作为自变量,行业减污降碳效果为因变量,采用赫克曼(Heckman)两阶段选择模型分析了数字经济对产业绿色化转型的影响。研究发现,产业数字化投入对碳排放和污染排放存在非线性影响。究其原因,是样本期内数字基础设施行业的数字化投入的边际收益小于边际成本,存在数字经济外部性的滞后效应[32]。农业则比较复杂,农业领域的研究认为数字技术对农业生产率有贡献,但这与资源禀赋的现存差异(如是否为粮食主产区、是否为旱地区域以及农户的收入)有关[33]。而服务行业因经营主体对信息技术的应用不足,导致信息通信技术(ICT)与经济效率之间联系不显著,比如旅游业[34]。在公共服务行业,警察部门由于自身的组织灵活性较低、无法根据技术进步及时动态调整自身的管理构架,使得信息技术对警察部门生产力(以犯罪率和警察部门清除犯罪的百分比作为衡量标准)的影响很小[35]。

近年来,学者将这一理论引申为解释具有“索洛悖论”现象即技术创新的“产生周期”的工具。孙毅等研究企业数字化转型对其绿色创新的影响时发现: 整体上,数字化技术对企业创新数量和质量具有推动作用,但创新激励效用的显现存在“形成周期”,即数字化转型存在“索洛悖论”。对此,可以从内部管理和融资约束两个方面进行理解: 一是新技术的应用需要相应的组织转型,这是一个旧的组织管理与新技术产生摩擦、不断调试并最终整合的过程。二是企业的数字化转型初期,初始投资远远高于其边际效益,数字化转型在企业业绩或增长方面直接的盈利效应尚未充分显现,由于信息不对称性,金融机构通常不愿意为其提供高额的融资服务。同时数字化转型为企业带来了额外的融资需求,挤出了企业有限的资金,导致更高的利息成本,加大了企业的融资约束。[36] 上述研究结果说明,一项技术创新带来正的外部性显现需要经历较长时间的滞后期,内在逻辑在于: 除了技术创新本身存在周期外,创新还需要与之匹配的孵化条件。[37-38] 杨帆和王满仓从研发投入扩大技术前沿差距,表现出“索洛悖论”特征这一现象出发,以中国信息技术产业上市公司为研究对象,得出了相似的结论: 研发投入需要与之匹配的人力资本配套才能实现对前沿技术的追赶,研发投入与人才的错位会导致技术追赶效应具有时滞性。[39] 直到今天,“索洛悖论”依然是学者关注和争论的焦点,虽然各行业特征不同,结果呈现部分差异,但基本达成一个新的共识,即创新技术转化成生产率等正的外部性需要组织形态、生产关系等配合。

技术进步将推动人工智能伦理的发展[40],基于已有文献与人工智能伦理发展现实条件的考量,我们认为 ChatGPT 对人工智能伦理推动也应该同样具有“技术创新伦理周期”特性,即生成式人工智能对社会伦理的促进作用也会具有“U 型”特征。当前隐私暴露带来的信任危机、算法不成熟导致的片面信息传播风险属于应用初期客观上需要面临的阶段性特征。接下来参考现有研究,以“索洛悖论”为分析工具,ChatGPT 为自变量,伦理发展为被解释变量,将伦理进步看作是科技创新投入的增函数,对上述问题做进一步分析。

(二) ChatGPT 的伦理“索洛悖论”

1. 以 ChatGPT 为代表的生成式人工智能与以往的技术创新的异同

历史上出现过多次重要的技术变革,进入 21 世纪以来人工智能技术得到蓬勃发展。多数学者认为,当今世界已进入了全新的发展阶段,并称之为工业 4.0。与以往的工业革命不同,人工智能技术可以给机器赋予智能,在许多方面替代人类劳动,可能会改变人类和机器在生产生活中的角色,如有的餐厅开始利用传菜机器人为顾客送菜。很明显,人工智能将在更大范围替代人工劳动。2022 年以来,ChatGPT 的出现标志着生成式人工智能时代的来临,目前它在许多行业已经被证明是有巨大价值的工具,有望掀起新一轮的革命。其能够实现低成本甚至零成本的自动化内容创建,彻底改变了众多行业的生产范式。ChatGPT 具有使用便捷,技术投入不确定性和长期性,知识边界管理、伦理风险规避等方面的可预测性较低,满足颠覆性技术创新的典型特征,因此,ChatGPT 将在总体层面和不可预测的未来发展上产生跳跃式革命[1]。但同时,从技术进步的本质看,ChatGPT 的发展是机器代替人类劳动传统的延续。

2. 以 ChatGPT 为代表的生成式人工智能伦理“索洛悖论”

发电机和蒸汽机从出现到其技术创新效应的显现分别经历了 40 年和 20 年,信息通讯技术也不例外,不过时滞期大大缩短,约为 10 年时间[41]。这种现象可以称之为技术迟延,时滞的存在说明技术扩散及其经济效应显现需要合适的框架条件。除了内在技术成熟度,它还受到外在信息技术基础设施、人力资本、管理实践及配套的监管机制等约束[25] [42]。

那么,ChatGPT 发展是否同样遵循“索洛悖论”?作为经济增长的众多因素之一,ChatGPT 符合“通用目的技术”特征[43],将成为推动经济发展的变革性力量。但同时,其作用的发挥不是无条件的。一是实现大型模型的稳定训练和获得优异的性能需要极高的计算成本:GPT-3 训练 1 次的费用是 460 万美元,总训练成本达 1 200 万美元,GPT-4 的训练成本则是 GPT-3 的 5 倍以上。二是稳定和持久的模型培训离不开微软 Azure 云平台的完整性和稳定性。三是大规模数据和大型模型训练需要持续的数据、代码和工程调优,这就需要员工具备丰富的系统优化经验。四是生成式人工智能进步需要监管机制及时补位和动态调整以降低潜在的“副作用”,平衡科技创新和社会发展。因此,根据技术创新的内在技术属性和外在孵化环境,可以合理猜测,ChatGPT 同样会陷入“索洛悖论”。相应地,其引起的伦理进步效应也会像其他技术进步一样呈现“索洛悖论”特征。

(三) 以 ChatGPT 为代表的生成式人工智能技术创新伦理周期

生成式人工智能发展有助于社会伦理进步[40],但其进步效应并不是一蹴而就的,从前期应用场景的拓展到中期社会、文化、伦理与技术的有机融合,再到后期产生实效,需要经历一定的过程。根据目标和任务不同,整个作用过程可以划分为商业分析、数据工程、机器学习建模、模型部署以及运作和监控 5 个发展步骤,伦理风险问题往往与人工智能生命周期的某个阶段相联系[44],因而技术创新推动伦理进步存在“技术创新伦理周期”。

生成式人工智能应用初期,受技术手段等限制,用户可以利用技术漏洞和信息不对称获取不当利益,如利用 ChatGPT 作弊通过资格考试,生成虚假热点新闻获取流量,操纵和散布带有政治立场或价值取向的观点制造冲突[16] [45]。低门槛和巨大的投机空间诱惑用户游走于伦理道德边缘,冲击了当前的社会伦理体系。在应用中后期,技术(或机器)的革新通过不断试错的动态方法促进知识体系不断突破局限[46]。同时,通过与 ChatGPT 的互动和模拟不同情景下的道德决策,监管机构、伦理专家和社会公众动态评估系统对于伦理问题的反应和表现,并提出批评、建议或完善监督机制,以确保技术发展与人类价值和社会利益一致,最终提高公众对人工智能技术和人类角色的认知,促进整个社会伦理

素养的进步[40]。因此，生成式人工智能对社会伦理的影响可能呈现出先下降至拐点后上升的“U 型”特征，需要跨过阵痛期才能带来社会伦理水平的提升。

根据伦理问题与人工智能生命周期各阶段的映射关系，伦理问题往往是由技术革新的某一阶段的某种原因引起的（如技术漏洞引发的投机行为）。因而，从搜索引擎到底层设计，人工智能技术的起飞往往伴随着伦理因素的考量[47]。就 ChatGPT 目前的功能实现和特征看，一方面，迭代和升级过程已经具备了超大的参数规模。ChatGPT-3 基于 8 000 亿个单词的语料库，包含了 1 750 亿个参数，最新版本的 GPT-4.0 的参数量级是 GPT-3.5 的 10 倍。模型扩容和预训练数据增加有助于捕获更复杂的语言模式和关系，保证功能稳定和可持续。另一方面，ChatGPT 以变换器（Transformer）作为通用模块接口，支持多模态输入与输出。它可以接收音频、图像等多种格式的输入，并形成相应的输出。还使用了一种后训练对齐的方法，通过与人类专家进行交互，这意味其具备处理更复杂和丰富的信息、提供更加符合事实和满足用户期望的能力。整体而言，ChatGPT 初步实现了人机交互质量提升、算力可靠和稳定。在此过程中，许多组织（包括人工智能公司、学术界和政府）也积极寻求解决人工智能伦理问题的可能框架、指南和原则，致力于为缓解当下人工智能应用引发的伦理风险问题提供有价值的指导。包括在生成式人工智能中嵌入伦理道德，持续推进法律法规以规范人工智能的开发和应用，使其在伦理理论框架下推理、决策和开发新技术，提高产品透明度和可解释性、减少机器学习的偏见或歧视[44]。面对当前阶段层出不穷的伦理风险问题，开放人工智能（OpenAI）公司也先后推出了漏洞赏金计划①和红队军团成员招募计划（Red Teaming Network）②，鼓励开展多领域的沟通对话，听取来自计算机和工程领域以外的，如政治学、经济学、社会学、心理学等领域的专家建议，致力于实现内部测试和外部监督结合，保障系统安全、可靠和客观公正，降低伦理风险问题发生的可能。

可以看出，伦理进步往往落后于技术创新，技术革新的“索洛悖论”特征导致其推动的伦理进步存在“技术创新伦理周期”。在过渡期内，技术处于逐步完善和迭代阶段，系统性约束开发与应用主体行为的伦理规范体系正逐步具体和清晰化，部分个体利用技术和监管漏洞实施不道德行为使得伦理问题暴露出来。因而当前涌现的伦理问题是 ChatGPT 技术生命周期内发展的阶段性产物，是技术自我修正与人工智能伦理调整和补位过程的正常现象。

四、应对生成式人工智能的技术创新伦理周期的政策建议

基于 ChatGPT 的分析表明，技术创新的“U 型”特征决定了公众科学素养的提高和相应监督机制的完善存在技术创新伦理周期。当前生成式人工智能应用过程暴露的伦理风险是技术创新伦理周期的阶段性特征，是技术创新周期与系统性规范体系动态演进过程的正常现象。随着生成式人工智能应用场景的不断成熟和法律制度的完善，新一代人工智能的广泛应用最终会促进社会伦理文明进步。为此，在周期内，需要对开发与应用主体加以约束和引导，保证其发展与应用符合伦理道德和社会价值，跨越生成式人工智能发展的“伦理风险鸿沟”。

2023 年 7 月 13 日，国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局公布了《生成式人工智能服务管理暂行办法》（以下简称《办法》），《办法》将从 2023 年 8 月 15 日起实施。《办法》强调鼓励人工智能创新发展，这表明政府发布《办法》不是限制生成式人工智能的发展，而是支持并从政策上保障其发展。《办法》针对生成式人工智能开发与应用的社会公德和伦

① 见 https://openai.com/blog/our-approach-to-ai-safety。

② 见 https://openai.com/blog/red-teaming-network。

理道德提出以下要求:坚持社会主义核心价值观;有效防范民族、信仰、国别、地域、性别、年龄、职业、健康等歧视问题的产生;有效防范不当竞争和垄断;保障个人合法权益;保证生成式人工智能服务透明,生成内容准确和可靠。结合索洛悖论的产生原因、阶段性特征和《办法》的基本原则,提出跨越生成式人工智能应用的"伦理风险鸿沟"的建议。

(一)推动构建生成式人工智能治理体系

生成式人工智能应用初期,生成式人工智能底层技术的可靠性和稳定性等方面的成熟度以及用户的人工智能素养亟待进一步提高,不法分子利用信息不对称实施不道德行为非法牟利。利用 ChatGPT 实施恶意造谣蹭热度、散布非法、虚假信息谋取不正当收益等,造成社会伦理水平的整体下滑。政府部门和生成式人工智能提供方可以通过营造支持生成式人工智能应用的社会环境和健全监管生成式人工智能应用的服务平台降低社会公众素养下降的可能。

一是政府部门应完善法律框架和健全监管体制机制,建立生成式人工智能伦理框架。明确生成式人工智能服务提供方和用户主体责任以及生成式人工智能在各种场景下的行为规范,对相关平台和用户追究相应的法律责任,确保生成式人工智能技术的研究和应用与道德原则相符。二是平台服务方应吸纳多元化开发主体,减少主观偏见的可能。生成式人工智能服务提供方可以考虑吸收不同文化背景、学科背景、不同地域的群体,在开发、人机互动应用和维护过程听取更多领域的意见,提高生成产品的多元价值观[48],减少文化背景引起的系统性偏见[47]。三是加强数据安全保护,保障用户隐私与安全。服务提供方应采取必要的技术措施,确保用户的个人信息不被非法获取、篡改、泄露;同时明确隐私政策,告知用户个人信息的收集、使用、存储、保护等情况,在明确告知用户的情况下使用,不得将用户的个人信息用于其他目的;加强用户授权管理,确保用户授权的范围和期限明确,不得超出用户的授权范围使用个人信息;建立健全监管机制,对违反隐私保护规定的行为进行追责,防范产品非法留存推断用户身份,杜绝产品根据用户输入信息对用户做画像,防止向他人提供用户输入信息,保护用户的合法权益。四是规范标识训练数据来源,提高数据透明度。服务提供方利用全球公开数据开展预训练和优化训练,应标明数据和基础模型来源。一是有利于减少知识产权纠纷,保证数据的合法合规;二是有助于结果的追溯和检验,提高数据的透明度和可靠性,减少生成结果不当引用;三是提高了虚假信息成本,降低由于模型训练生成虚假信息的可能。

(二)规范监管保障生成式人工智能安全发展

在生成式人工智能应用中期,生成式人工智能应用的约束性规范基本形成,技术漏洞陆续得以修复,社会伦理的基线水平得以保证。生成式人工智能提供方应加强生成式人工智能服务平台监管,保证生成的内容合法合规、尊重社会公德和伦理道德、提升服务透明度,提高生成内容真实性和可靠性。

一是提高标记人员素养,规范标记规则,提高训练数据质量。服务提供方应当制定清晰、具体、可操作的标注规则,对标注人员进行必要培训,提高业务能力,防范由于人工操作不当生成虚假信息的风险;监督指导标注人员规范开展标注工作,开展数据标注质量评估,增强训练数据真实性、准确性、客观性和多样性。二是平台应建立和完善安全事件应急预案和响应机制,提高及时处理和恢复能力。具体包括:其一,加强对大模型预训练语料的数据监管,识别潜在的恶意用户,预测和干预可能实施的危险行为,限制其对生成式人工智能的访问权限与范围,并规范其使用场景与使用方法,降低恶意用户攻击的可能和范围。其二,加强安全防护技术的使用和监管,将 ChatGPT 与入侵检测系统、网络监控等工具结合,提高其更准确和快速的识别和报告潜在犯罪或不道德行为的可能。同时,利用 ChatGPT 的生成能力,打造攻击模拟现场,并反复测试防御系统的稳定性和可靠性。其三,不断优化和迭代安全模型,以保持其有效性。确保攻击发生后,ChatGPT 能够迅速响应,自动执行一些基本的响应操

作,如隔离被攻击的系统,阻止攻击者的进一步渗透等,并对攻击事件进行分析,发现和修补漏洞,提出可能的解决方案,提高事件响应的速度和效率,减少潜在的损失。

(三) 优化落实生成式人工智能可持续发展环境

生成式人工智能应用后期,市场规模基本稳定。政府部门和生成式人工智能服务提供方应致力于提供更完善的配套设施、培养更多的高水平的研发和应用人才和探索更多(交叉)领域的实践场景,提高公众的人工智能素养,促进社会伦理整体水平的提高。

一是政府部门应继续加大基础设施建设。推动落实数字化基础设施和公共训练数据资源平台建设,提升现有技术设施智能化水平和覆盖范围,提升基础设施对技术创新的支撑能力,拓宽公众科学素养提高渠道。二是政府应考虑从顶层设计入手,培养和吸引技术人才,破解"卡脖子"技术难题。在《办法》鼓励生成式人工智能技术在各行业、各领域创新应用的背景下,着重培养和吸引人工智能算法、软件、芯片等技术人才。三是加强生成式人工智能知识的拓展和教育,提高大众生成式人工智能素养。政府部门和服务提供方应关注生成式人工智能在各领域的应用和推广,并不断完善和应用生成式人工智能技术,同时加大生成式人工智能宣传和普及,引导公众正确认识和规范使用生成式人工智能,适应生成式人工智能带来的新变化。

参 考 文 献

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

[2] 何大安,许一帆. 人工智能应用扩张的经济学分析——兼谈 ChatGPT 对厂商经营活动的影响[J]. 社会科学战线,2023(9).

[3] 陈锐,江奕辉. 生成式 AI 的治理研究:以 ChatGPT 为例[J]. 科学学研究,2023(10).

[4] 陈元,黄秋生. ChatGPT 技术中的人工智能伦理风险及其科学祛魅[J]. 湖南科技大学学报:社会科学版,2023(3).

[5] 冯涛,董嘉昌,李佳霖. ChatGPT 等生成式人工智能对我国经济高质量发展的双重影响及其应对[J]. 陕西师范大学学报:哲学社会科学版,2023(4).

[6] PLOYHART R E, BARTUNEK J M. Editors' Comments: There is Nothing so Theoretical as Good Practice——A Call for Phenomenal Theory[J]. Academy of Management Review, 2019(3).

[7] SAETRE A S, VAN DE VEN A. Generating Theory by Abduction[J]. Academy of Management Review, 2021(4).

[8] GRAEBNER M E, KNOTT A M, LIEBERMAN M B, et al. Empirical Inquiry Without Hypotheses: A Question-driven, Phenomenon - based Approach to Strategic Management Research[J]. Strategic Management Journal, 2023(1).

[9] DWIVEDI Y K, KSHETRI N, HUGHES L, et al. So What if ChatGPT Wrote It? Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy[J]. International Journal of Information Management, 2023(10).

[10] 孔德臣,姜迎春. ChatGPT 等新一代人工智能的特性及其数字经济效应——基于马克思的机器与异化理论[J]. 经济问题,2023(7).

[11] 邱冬阳,蓝宇. ChatGPT 给金融行业带来的机遇、挑战及问题[J]. 西南金融,2023(6).

[12] ZHU C J, SUN M, LUO J T, et al. How to Harness the Potential of ChatGPT in Education?[J]. Knowledge Management & E-Learning-an International Journal, 2023(2).

[13] 李勇坚. ChatGPT 与经济增长:影响机制与政策框架[J]. 新疆师范大学学报:哲学社会科学版,2024(2).

[14] NOY S, ZHANG W. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence[J]. Science, 2023(6654).

[15] WU T Y, HE S Z, LIU J P, et al. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development[J]. Ieee - Caa Journal of Automatica Sinica, 2023(5).

[16] STOKEL-WALKER C, VAN NOORDEN R. The Promise and Peril of Generative AI[J]. Nature, 2023(7947).

[17] MOONS P, VAN BULCK L. ChatGPT: Can Artificial Intelligence Language Models Be of Value for Cardiovascular Nurses and Allied Health Professionals [J]. European Journal of Cardiovascular Nursing, 2023(7).
[18] 于水,范德志.新一代人工智能(ChatGPT)的主要特征、社会风险及其治理路径[J].大连理工大学学报:社会科学版,2023(5).
[19] 蔡镇疆,车宇璐.ChatGPT 的法律人格认定与生成内容性质探讨[J].新疆师范大学学报:哲学社会科学版,2024(2).
[20] MARCHANDOT B, MATSUSHITA K, CARMONA A, et al. ChatGPT: The Next Frontier in Academic Writing for Cardiologists or a Pandora's Box of Ethical Dilemmas [J]. European Heart Journal Open, 2023(2).
[21] LUBOWITZ J H. ChatGPT, An Artificial Intelligence Chatbot, Is Impacting Medical Literature [J]. Arthroscopy, 2023(5).
[22] 荆林波,冯永晟.信息通讯技术、生产率悖论与各国经济增长[J].经济学动态,2010(6).
[23] 陈欢,周密,闫文凯."索洛悖论"的存在性研究——基于江苏省地级市的经验证据[J].软科学,2018(4).
[24] 贾伟,王丽明,毛学峰,等.中国农业企业存在""出口—生产率悖论"吗?[J].中国农村经济,2018(3).
[25] 杜传忠,郭美晨.信息技术生产率悖论评析[J].经济学动态,2016(4).
[26] 何小钢,梁权熙,王善骝.信息技术、劳动力结构与企业生产率——破解"信息技术生产率悖论"之谜[J].管理世界,2019(9).
[27] DAVID P A. The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox [J]. The American Economic Review, 1989(2).
[28] BRYNJOLFSSON E, HITT L. Paradox lost? Firm-Level Evidence on the Returns to Information Systems Spending [J]. Management Science 1996(4).
[29] TRIPLETT J E. The Solow Productivity Paradox: What Do Computers Do to Productivity? [J]. The Canadian Journal of Economics, 1999(2).
[30] 陈楠,蔡跃洲.人工智能、承接能力与中国经济增长——新"索洛悖论"和基于 AI 专利的实证分析[J].经济学动态,2022(11).
[31] WEN H W, JIANG M, ZHENG S F. Impact of Information and Communication Technologies on Corporate Energy Intensity: Evidence from Cross-country Micro Data [J]. Journal of Environmental Planning and Management, 2022(10).
[32] 方颖,余兴锦.产业数字化的减污与降碳效应——基于"生产率悖论"的研究视角[J].系统工程理论与实践,2023(8).
[33] 张永奇,庄天慧,杨浩.数字经济下农业生产率悖论与破解之道研究——基于 CLDS 小农户数据的经验考察[J].西南民族大学学报:人文社会科学版,2023(7).
[34] 于婷婷,左冰.信息化对旅游经济效率的影响及其作用机制研究[J].地理科学,2022(10).
[35] GARICANO L, HEATON P. Information Technology, Organization, and Productivity in the Public Sector: Evidence from Police Departments [J]. Journal of Labor Economics, 2010(1).
[36] SUN Y, HE M Y. Does Digital Transformation Promote Green Innovation? A Micro-level Perspective on the Solow Paradox [J]. Frontiers in Environmental Science, 2023(11).
[37] 李静,楠玉,刘霞辉.中国研发投入的"索洛悖论"——解释及人力资本匹配含义[J].经济学家,2017(1).
[38] 邱煜,潘攀.风险投资能打破创新的"索洛悖论"吗?[J].经济经纬,2019(6).
[39] 杨帆,王满仓.研发投入与技术前沿差距的"索洛悖论"——基于研发人力资本的解释[J].现代财经(天津财经大学学报),2020(12).
[40] STAHL B C, EKE D. The Ethics of ChatGPT—Exploring the Ethical Issues of An Emerging Technology [J]. International Journal of Information Management, 2023(10).
[41] BRYNJOLFSSON E, ROCK D, SYVERSON C. Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics [R]. NBER Working Paper No 24001, 2019, 23-57.
[42] 何小钢,王善骝.信息技术生产率悖论:理论演进与跨越路径[J].经济学家,2020(7).

[43] 王俊秀. ChatGPT 与人工智能时代: 突破、风险与治理 [J]. 东北师大学报: 哲学社会科学版, 2023(4).
[44] HUANG C, ZHANG Z, MAO B, et al. An Overview of Artificial Intelligence Ethics [J]. IEEE Transactions on Artificial Intelligence, 2022(3).
[45] STOKEL - WALKER C. Chatgpt Listed as Author on Research Papers [J]. Nature, 2023(7945).
[46] 任晓明, 林艺霏. 人工智能视野下的知识体系修正理论 [J]. 陕西师范大学学报: 哲学社会科学版, 2022(2).
[47] BORENSTEIN J, GRODZINSKY F S, HOWARD A, et al. Al Ethics: A Long History and a Recent Burst of Attention [J]. Computer, 2021(1).
[48] 张钺, 李正风, 潜伟. 从 ChatGPT 到人机协作的知识共建 [J]. 科学学研究, 2023(12).

[责任编辑　蒋万胜]

The Solow Paradox of Ethical Risk in Generative Artificial Intelligence ——A Case Study of ChatGPT

LEI Hong-zhen, LIU Chao, LAN Juan-li
(*School of International Business, Shaanxi Normal University, Xi'an* 710019, *Shaanxi*)

Abstract: Generative Artificial Intelligence(AI) while driving breakthrough applications in technological innovation, also carries many social and ethical risks: privacy exposure and information security control risks; the risk of information spread caused by the derivation of false information; the decision-making and policy risks associated with algorithmic bias, as well as the risk of vague definitions of liability. How to view and respond to these risks, this paper takes the application of ChatGPT as an example, introduces the "Solow Paradox" theory, regarding ethical progress as an increasing function of technological innovation investment, and proposes that ChatGPT has a "technological innovation Solow cycle" characteristic in promoting ethics. Countermeasures and recommendations have been proposed to bridge the ethical risk gap in generative AI: we should establish and improve generative AI laws and regulations, and promote the establishment of a governance system for generative AI. The generative AI supervision platform needs to be strengthened and the process standardization management needs to be improved to standardize and ensure the safe development of the generative AI. The sustainable development environment of the generative AI needs to be further implemented and optimized, and the possibility of reshaping more industries should also be explored.

Key Words: generative Artificial Intelligence; technological innovation; Solow paradox; ethical risks