文章编号：0253-374X(2023)05-0652-08
DOI：10. 11908/j. issn. 0253-374x. 23054

基于LDA模型的人工智能伦理准则体系研究

贾　婷[1,2]，陈　强[1]，沈天添[1]
（1. 同济大学 经济与管理学院，上海 200092；2. 大理大学 经济与管理学院，云南 大理 671000）

摘要：利用隐含狄利克雷分布(Latent Dirichlet Allocation，LDA)模型对全球人工智能伦理准则体系进行内容分析，通过对准则关键词的阶段差异对比、核心主题的聚合体系分析，总结出世界范围内人工智能治理共识的6大演变趋势，均表明实践导向已成为核心共识。但人工智能伦理准则存在“可操作性受限、考量要素不同质、伦理洗白、法律转化难”等实践适用性困境，亟须在技术、组织及制度层面弥补实践缺口。

关键词：科技伦理风险；人工智能伦理准则；演进趋势；实践困境
中图分类号：C935　　**文献标志码**：A

Latent Dirichlet Allocation Model-based Research on System of Ethical Guidelines for Artificial Intelligence

JIA Ting[1,2], *CHEN Qiang*[1], *SHEN TianTian*[1]
(1. School of Economic and Management, Tongji University, Shanghai 200092, China; 2. School of Economic and Management, Dali University, Dali 67100, China)

Abstract: Based on the latent Dirichlet Allocation (LDA) model, an analysis was made of the content of the global artificial intelligence ethical guidelines system. Six major trends of the worldwide consensus on artificial intelligence governance were summarized by comparing the stage differences of keywords and aggregating the core themes of the guidelines. Results indicate that practice orientation becomes the core consensus. However, the implement of the artificial intelligence ethical guidelines faces the dilemmas such as limited operability, heterogeneous considerations, ethical whitewashing, and difficult legal translation. Fianally, a discussion was made on the details to bridge the practical gap.

Key words: ethical risks in technology; artificial intelligence ethical guidelines; evolutionary trends; practical dilemmas

以科技伦理准则为代表的“软体系”先行，在全球范围内广泛讨论技术风险、提出预期治理目标、构建预测性治理机制、细化敏捷性治理方案，将是实现硬法规制的先导性探索。以人工智能技术为例，近年来在全球范围内由政府、企业、社会机构、国际组织、学术团体等利益相关主体提出的治理原则或倡议已有150多个。学者们普遍认为人工智能伦理准则要求下的新兴治理技术的发展是实现可信AI实践的关键，但会涉及复杂的、差异化的技术环节和周期[1]，导致现有的准则规范并不能影响利益相关体在参与AI开发和应用过程中的合规行为[2]。相当多的准则规范是由私人部门所提出，其可能作为抵制政府强监管的借口而最终流于形式[3]，还有学者指出AI伦理准则在内容上存在缺失或争议[4]。为探究科技伦理准则有效落地实施的提升路径，本文以人工智能伦理治理为切入点，通过分析现阶段人工智能伦理准则体系的内容及其演变趋势，进而探究准则实施的困境及破解之道，尝试回应这些问题。

1　人工智能伦理准则体系的内容分析

以代表性人工智能准则文本为研究对象，采用内容分析法，利用文本挖掘手段提取准则文本中的关键词和主题信息，聚合出核心主题并分析其内容特性、主体特征及影响范畴，并以此为基础探究准则间的内在联系。同时，通过分析AI准则关键词的阶段性特征，研判全球AI伦理共识的演变趋势，以发

收稿日期：2023-02-22
基金项目：国家社会科学基金重大项目(21ZDA018)
第一作者：贾婷(1984—)，女，讲师，博士生，主要研究方向为创新与技术管理、科技伦理治理。E-mail：jttj@tongji. edu. cn。
通信作者：陈强(1969—)，男，教授，博士生导师，管理学博士，主要研究方向为科技创新治理。E-mail：chenqiang@tongji. edu. cn

论文拓展介绍

现AI伦理治理的关键走向,为进一步探究伦理准则的实践应用提供理论分析基础。

1.1　数据来源与研究设计

1.1.1　数据来源

依托中国科学院自动化研究所“类脑认知智能实验室和人工智能伦理与治理研究中心”开发的LAIP——链接人工智能准则平台。选取该数据库中收录的92个2016—2022年之间发布并被明确逐项记录的代表性AI伦理准则,以此为数据样本来进行相关内容分析,准则基本情况统计如表1所示。

表1　准则基本情况统计

Tab.1　Basic statistics on guidelines

发布主体类型	准则数量
政府、政府间组织	35
学术界、非营利组织、非政府组织	30
行业组织	27
国家及区域(前5位)	**准则数量**
美国	27
国际组织	15
中国	12
英国	7
日本	5
加拿大	5
发布时间阶段	**准则数量**
2016－2017	18
2018－2019	56
2020－2022	18

1.1.2　研究方法与研究过程

研究基于Python语言构建整体模型对准则文本进行量化分析。一方面,通过抽取阶段关键词汇挖掘全球范围内人工智能伦理准则在不同阶段的关注重点,并绘制可视化词云图,更直观地判断阶段差异;另一方面,提取准则主题,进一步理解人工智能伦理准则的内涵分布及演进方向,具体研究过程如图1所示。

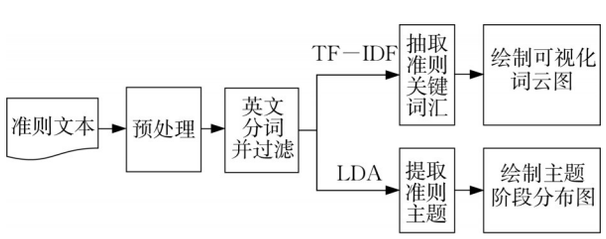

图1　研究过程

Fig.1　Research process

词频逆文档频率TF-IDF(term frequency-inverse document frequency)算法,在文本挖掘中广泛应用于衡量词汇对文本的重要程度。该方法综合考虑词汇的频数和在文本中的重要程度,提取具有显著差异的关键词汇,以更好地代表准则文本中的信息。特定词汇的重要性与其在该文本中的出现次数成正比,与其在所有文本中的出现次数成反比。本文采用平滑后的IDF计算方法,词汇W在一篇文本中的TF-IDF权重计算为

$$\text{TF-IDF}(W)=\frac{N_{\text{w}}}{N}\times\left(\log\left(\frac{Y+1}{Y_{\text{w}}+1}\right)+1\right)$$

式中:N_{w}表示词汇w在该文本中出现的次数;N表示该文本总词汇数;Y表示总文本数;Y_{w}表示含有词汇W的文本数。对TF-IDF权重进行归一化处理后,每篇文本词汇权重之和等于1。计算候选词汇在准则文本中的TF-IDF权重,得到权重较高的多个名词性质词汇,作为该准则文本的关键词汇,反映不同时期人工智能伦理准则着力点的变化过程。

隐含狄利克雷分布(Latent Dirichlet Allocation,LDA)模型,基于Dirichlet分布和贝叶斯算法提取文本所包含的潜在主题信息和主题的词汇分布。该模型考虑了主题概率分布的先验知识,认为准则内容围绕一个或多个主题展开,而主题由词汇的概率分布确定。对于某篇文本中的每一个词汇,需要从该文本的主题分布中随机选择一个主题,再从主题对应的词汇分布中随机选择一个词汇,重复该操作直至所有文档全部生成,最终求解获得准则文本的主题分布和主题的词汇分布,其原理如图2所示。图中,M表示文本数量;N_m表示文档中的词汇;K表示主题数量;α为主题分布的先验分布,θ_m为第m篇文本的主题分布,$z_{m,n}$为第m篇文本中第n个词对应的主题,$w_{m,n}$为第m篇文本中的第n个词;β为词汇分布的先验分布,φ_k为第k个主题的词汇分布。通过评估不同主题数模型的困惑度来确定最优的模型主题数量,从而提取准则主题及其词汇分布,并在此基础上理解人工智能伦理准则的内涵及发展历程。

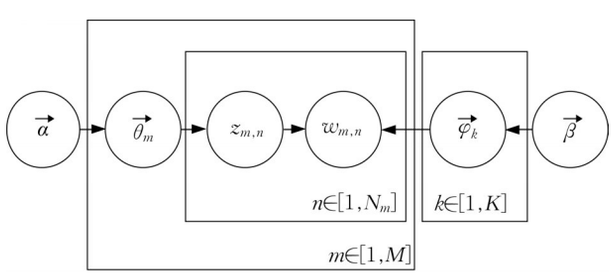

图2　LDA模型

Fig.2　LDA Model

1.2 研究结果分析

1.2.1 以准则关键词对比阶段差异

图3统计了2016—2022年间全球范围内发布的AI伦理准则中11个核心关键词出现的频次[5],展现了全球视角下对人工智能伦理风险关注的重要方向和主要层面。有学者指出,不同国家和地区出台的人工智能伦理准则的关注重点存在差异,如中国更关注人类福祉,要求尽可能降低对用户造成的消极后果;欧盟更关注公平,限制性原则较多;美国更关注可控性,较少关注分享[6],这在一定意义上表明了跨国伦理挑战的存在。

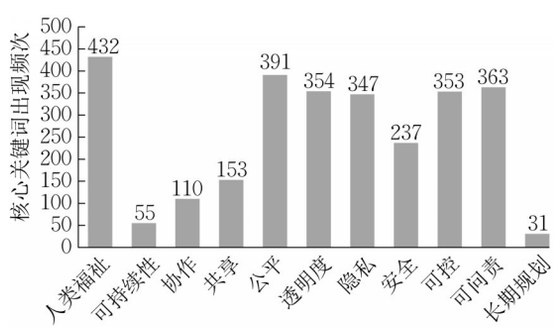

图3 AI伦理准则关键词统计(2016—2022)

Fig.3 Keywords statistics on AI ethical guidelines (2016—2022)

结合人工智能技术发展轨迹和本文搜集的数据特征,拟从初步探索(2016—2017),原则爆发(2018—2019)、实践应用(2020—2022)3个阶段来分析人工智能伦理准则的阶段性特征,如图4所示。

图4 阶段AI准则关键词分析

Fig.4 Keywords analysis of each phase

从AI伦理风险议题被提出到各界纷纷提出或制定伦理准则并广泛寻求共识,再到探索把准则转化为实践的机制、做法、工具等,每个阶段的治理侧重点和准则表述存在差异。

首先,AI准则的制定密切围绕技术实体发展,“技术、算法、系统、数据”等关键词在3个阶段的权重都比较高,但经历了原则大爆发阶段后,准则逐步跳出技术系统,开始关注部署应用和社会影响,对风险的关注度逐渐提升。其次,进入实践应用阶段(2020—2022),考量伦理准则在实践中应用的权重在逐步提升,application(应用)在词云图中被重点凸显出来;从实践参与主体的角度来看,第1、2阶段较为笼统的谈及利益相关者作用的发挥,在第3阶段则强调actor(行动者)的实践行为;再次,在AI伦理准则发展的3个阶段中,法律一直没有被凸显出来,这意味着无论是从行为规制角度还是行为奖励的角度,都鲜少有将人工智能伦理规约确认为法律规范的实质性探讨。

1.2.2 以准则主题考察体系演变

通过对92个准则文本提取准则主题,参照LDA模型的困惑度曲线和主题含义确定准则核心主题的数量为5,对词汇内容进行概念化后归纳的主题类别名称为“通用性准则”、“行业发展准则”、“技术系统准则”、“制度干预准则”、“实践性准则”。各核心主题对应的部分代表性词汇如表2所示。

由表2可以发现,通用性准则聚类的关键词比较全面,既涉及到人类福祉、可持续发展等高阶价值层次,也聚焦到技术系统、数据安全、个体隐私、使用者权利等低阶价值层次,其发布主体分布均衡,体现了政府、业界及社会对AI伦理治理问题讨论的广度和宽度。行业发展准则更关注AI产业数字化赋能过程中的自律、合作、数据共享、技术安全等方面,其数量相对较少,以行业组织为发布主体。技术系统准则关注AI技术研发及应用中的透明性、可预测性及数据公平等,强调AI产品和服务安全,重视保障客户、政府及雇员等利益相关者的权益,以领军企业为发布主体。以政府及政府间组织为主体发布的制度干预准则,聚焦于AI技术研发及应用的可问责性、推进风险评估,寻求将伦理规范上升为法律规制的可能。实践性准则将负责任研发和创新作为发展宗旨,注重应用层面的实质性行动,也提出了具体的政策导向,其发布者主要以政府和政府间组织为主,对AI产品的伦理品质提出了要求。

本文也从时间轴角度对这5个核心主题的阶段化分布做了可视化呈现,如图5所示,其结果与上文所划分的3个准则发展阶段相吻合。尤其在2018—2019年这一阶段,呈现出“多管齐下”的态势,准则指向性较多,5类核心主题都有分布。值得注意的是,制度干预性准则的数量不多,关注度不高,这与目前被广泛讨论的议题相关,即在人工智能技术现阶段发展中进行制度干预的程度和方式该如何把握?也在一定程度上说明在人工智能伦理制度化过程中存在伦理治理的困境。

表2　准则核心主题

Tab.2　Core themes of the guidelines

核心主题	词项	相关度	词项	相关度	发布主体数量
通用性准则	系统	0.055 42	人类福祉	0.011 78	a(18);b(21);c(18)
	数据	0.027 79	发展	0.011 45	
	决策	0.015 54	人权	0.011 26	
	技术	0.015 51	使用权	0.010 94	
	原则	0.015 43	隐私	0.010 03	
行业发展准则	智能	0.048 40	技术	0.012 89	a(1);b(2);c(1)
	权利	0.024 43	安全	0.011 51	
	发展	0.018 09	条例	0.011 42	
	可循环	0.014 28	数据	0.010 80	
	保障	0.013 49	生命权	0.010 10	
技术系统准则	发展	0.017 64	规则	0.012 95	a(4);b(3);c(5)
	应用	0.016 52	服务	0.012 27	
	产品	0.014 90	客户	0.010 99	
	技术	0.013 21	数据	0.010 14	
	权利	0.013 20	安全	0.008 87	
制度干预准则	系统	0.039 63	研发人员	0.015 96	a(2);b(1);c(0)
	使用者	0.037 53	应用	0.014 98	
	数据	0.020 68	部署者	0.013 10	
	风险	0.020 34	准则	0.012 04	
	注意	0.018 70	判断	0.011 58	
	事项		责任		
实践性准则	系统	0.025 21	使用	0.013 16	a(10);b(3);c(1)
	开发	0.020 35	代码	0.011 12	
	行动者	0.020134	技术	0.011 04	
	儿童	0.016 52	机构	0.010 14	
	应用	0.014 54	方法	0.008 49	

注:a代表 政府和政府间组织; b代表Academia, NPO & NGO; c代表行业; a(18)代表由政府和政府间组织发布的通用性准则有18个

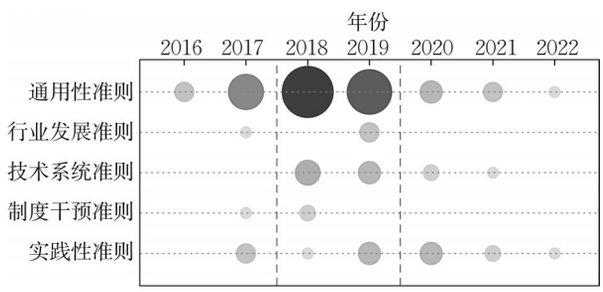

图5　核心主题的阶段化分布

Fig.5　Phased distribution of core themes

同时,本文选取准则发布数排名前五的国家(详见表1),在国别维度考察每一份准则的核心主题关联情况,如图6所示,灰度越深,代表该主题分布概率越高。整体来看,与技术发展程度及现阶段科技伦理规制的需求相适应,各国各类主体对核心主题的关注度各有侧重,围绕通用性准则相互补充,丰富了AI伦理治理议题讨论维度。但细化到具体国家层次,研究发现核心主题分布并不均衡,一是存在主题缺失,如中国和国际组织未出现“制度干预准则”主题,日本和美国未探及“行业发展准则”主题;二是主题分布单一,如德国发布的AI伦理准则主要对应到“技术系统准则”主题,俄罗斯大多围绕“实践性准则”主题设置内容,英国、加拿大则重点聚焦于通用性准则。

2　人工智能伦理治理共识的演变趋势

总体来看,全球范围内的代表性人工智能伦理准则一方面肯定了人工智能技术全领域应用给社会带来的巨大变革,另一方面也充分表达了对依托数据、算法和算力的AI技术应用的潜在伦理风险隐忧,并广泛寻求共识以实现规范发展。本文以同一年发布的准则中主题分布概率的算数平均值作为当年的主题强度[7],主题强度逐年演变趋势如图7所示。

从准则的具体内容来看,这些伦理治理共识的演变呈现出6大趋势(如图8所示):

(1)由理解风险走向解决风险。世界各大经济体逐渐意识到AI技术快速迭代对社会的重塑效应及潜在问题,应对风险的紧迫感日益增强。

(2)由强调技术安全转向技术可信。2018年之后,内涵更为全面的“可信”价值逐渐发展为人工智能技术研发的引领性共识,除强调人工智能技术要“安全可控”外,“透明可释、数据保护、明确责任、多元包容”等可信特征也被认可。

(3)由关注伦理价值判断走向实操应用。人工智能伦理治理进一步转向实践层面,探索一致性、兼容性价值规范落地的可能。

(4)伦理干预由设计阶段拓展到全周期。在联

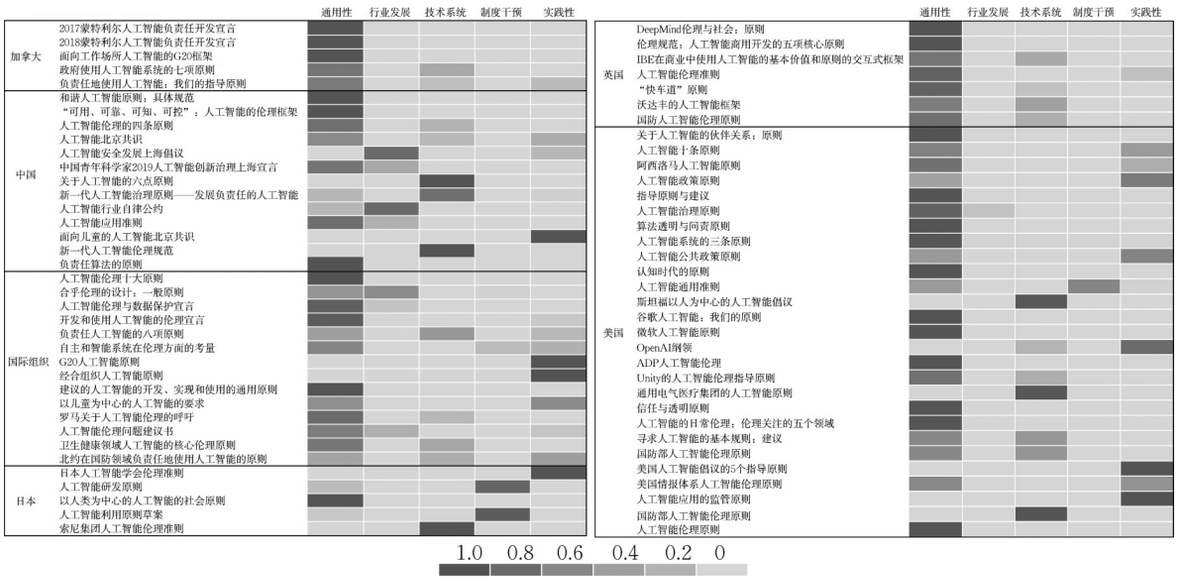

图6 核心主题分布国别比较

Fig.6 Country comparison of core theme distribution

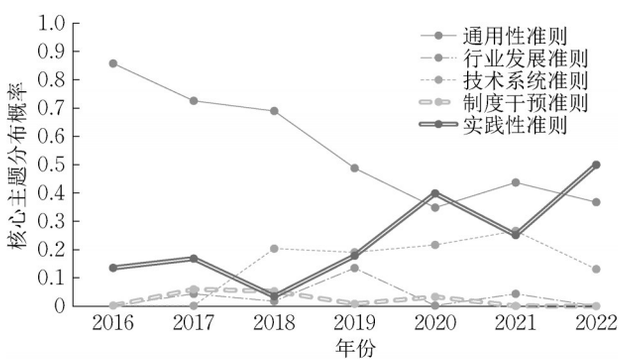

图7 主题强度逐年演变趋势

Fig.7 Year-on-year trends in the evolution of the intensity of the theme

合国教科文组织发布的《人工智能伦理问题建议书(2021)》中“AI系统的整个生命周期”出现频次多达51次,关于AI技术的治理阶段性的认识逐渐清晰和深入。

(5)治理主体由单一扩展到多元。相关领域的学术机构、国际组织、行业协会、大型科技公司与政府共同作为治理主体,探讨可信AI实践的可能。

(6)治理领域由宽泛趋于具体。从2020年开始,在对风险认知逐渐清晰的基础上,针对军用、情报、政府推广等具体领域的指引原则和具体标准开始实施,AI伦理治理的领域在不断聚焦。

综上所述,人工智能伦理治理共识从“风险认知、

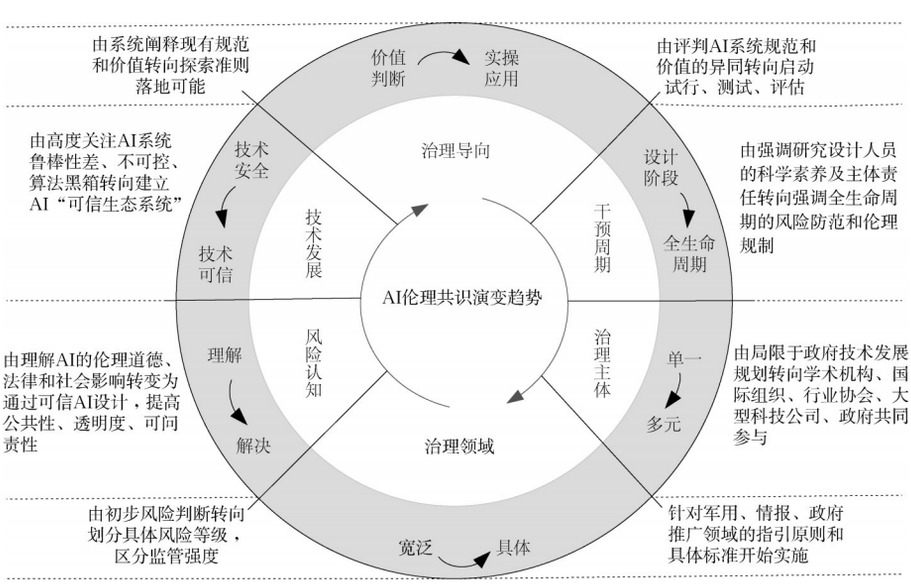

图8 AI伦理共识的演变趋势

Fig.8 Trends in the evolution of AI ethical consensus

技术发展、治理导向、干预周期、治理主体、治理领域”等多个维度都渐渐指向了“实践”。AI伦理准则虽已达到一定体量,也随着技术成熟度提升、应用场景具体化及风险认知科学化不断演变,并在全球范围内形成了若干关键共识,期望通过关键共识及规范要求的实施来合理控制风险。但是,近年来业界及学术界都指出了伦理准则落地难的问题,将伦理价值观纳入当前的AI治理框架还需克服诸多挑战。

3　人工智能伦理准则的实践困境

现阶段,关于人工智能伦理治理的讨论大多是通过关注关键准则来实现的,但也有学者指出AI伦理准则间存在冲突、重叠及缺失,归属领域或参与者不清晰,甚至存在实质性分歧[4],难以找到实践落点,需要进一步厘清关键核心原则间的关系。

结合上文聚合出的“通用性准则、行业发展准则、技术系统准则、制度干预准则、实践性准则”5类核心主题,本文试图从组织、技术和制度3个层面来分析关键核心原则的关系,并探究实践困境,如图9所示。

技术层面指向“技术系统准则”,结合关键词项相关度,选取“透明性、可预测性和公平性”作为核心原则。算法与数据的透明性既涉及AI技术本身、也涉及AI系统开发行为,可预测性则可看作是透明度的一个子集[8],透明可释的技术过程是预测智能行为的支点,促进“负责任”研发行为,并使“可问责”成为可能。此外,数据偏见及数字鸿沟等公平性问题也属于技术系统需要克服的难题,而可预测性是保证公平的基础。组织层面指向“实践性准则”和“行业发展准则”,选取“负责任”和“合作”作为核心原则。“负责任”侧重于AI研发行为要合乎伦理,“合作”是行业打破数据孤岛和技术壁垒,营造自律、持续发展氛围的有效机制;负责任研发行为有益于推动可预测性,保障公平性,促成组织间数据共享和模型开发等合作行为,形成良性发展的生态,而建立在合乎伦理要求前提下的合作也有利于推进负责任研发、推进可问责的实现。制度层面指向“制度干预准则”,选取“可问责”作为核心原则,旨在厘清利益相关者的责任问题,通过明确责任来实现责任共担,激发负责任行为。上述原则引导的最终目标是共同促进通用性准则的实现,以真正实现人类福祉、保证技术安全和可持续发展。关键伦理准则关系的厘清,是完善人工智能伦理治理框架的基础。

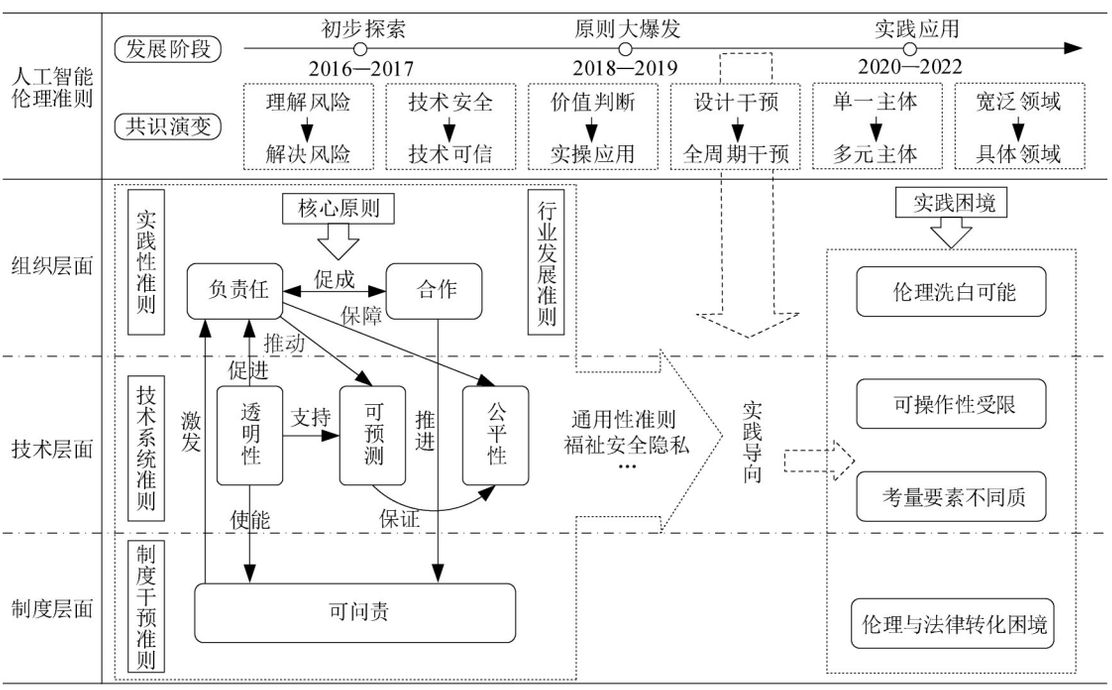

图9　AI核心伦理原则关系及实践困境

Fig.9　AI core ethical principle relationship and practice dilemma

而相关研究表明,现有的人工智能伦理准则并不能完全在实践中被采用[3,9],VAKKURI等人将39家与AI系统合作的公司的实践与“值得信赖的人工智能道德准则”中7个关键要求进行了比较,发现了人工智能伦理准则与实践之间确实存在显著差距,尤其针对软件开发的社会和环境福利需求以及多样

性、非歧视和公平的要求,并没有被充分考虑和解决[10]。基于前期探索,本文将人工智能准则的实践困境概括为以下几点:

3.1 伦理准则可操作性受限

人工智能伦理准则的出台为可信AI实践提供了探索的起点,但准则的可操作性受到多重因素影响。首先,伦理准则不是现成的方法,在实践中运用准则需要进行额外的工作,如开发关键领域的治理性技术和工具、形成组织内部的管理制度、提升实践能力等等,并要尽可能让伦理准则对技术开发者、部署应用者及产品使用者都更实用,而针对道德困境,通过直接设计数学函数为伦理规则和价值观建模很难[1];其次,准则往往关注的是AI系统设计和开发中较宏观的问题,而业界目前可用的、备受关注的治理工具仅能覆盖研发过程中的一小部分伦理风险,还不存在项目级方法[8]来提供更为便捷且普适性的治理思路和治理框架。再次,针对一些影响周期长、不确定的伦理问题,现阶段所提倡的伦理准则存在缺失或争议,作用边界模糊,技术端应对标的不明确。

3.2 伦理准则考量要素不同质

对于准则中提到的风险问题,很难评估现实努力在多大程度上实现了既定目标,或者是否存在相互矛盾的可能[11]。第一,伦理准则间关联交错难厘清,如“可预测性原则”与“透明性原则”难区分,“可问责性原则”与“负责任原则”的作用重点不同但表述相似,有的伦理原则作用于组织层面,有的伦理原则直接指向技术层面,但原则间又存在互动关系,容易产生错误引导;第二,伦理准则内涵理解存在差异,全球范围内的文化差异导致风险评估的思路和方法存在差别,技术标准不统一,伦理准则的规约范围不同,制度规范的力度和侧重点不同,不利于参与全球人工智能伦理治理合作。

3.3 伦理准则存在伦理洗白可能

企业是可信AI实践的主体,人工智能伦理准则要与企业的技术研发活动、产品开发及部署应用环节相对应,要与企业文化相一致,并在企业成员之间有效内化,伦理准则的实施还需要企业组织制度的保障。企业以盈利为导向的生存守则与为人类谋福祉的伦理准则之间存在逻辑冲突,使得私营部门在AI伦理领域的参与受到了质疑。企业参与制定伦理准则,可能仅是虚假信号,旨在推迟或完全避开监管[11],为AI技术系统和产品贴上“道德标签”,实现伦理“拔高”,但由于本身治理技术及能力缺乏,或者只是将其作为维护企业社会形象的工具,成为逃避监管的手段,存在伦理洗白的可能性。此外,目前由大型科技企业在技术、组织制度层面围绕伦理风险治理做出的探索,是否能辐射到智能转型中的大部分中小企业和以AI为核心的初创企业,成为产业生态的重要组成部分,也未有定论。

3.4 伦理准则与法律存在转化困境

两办印发的《关于加强科技伦理治理的意见》指出,‘十四五’期间,重点加强生命科学、医学、人工智能等领域的科技伦理立法研究,及时推动将重要的科技伦理规范上升为国家法律法规。这一意见聚焦新兴技术的发展对传统伦理观念和现行法律制度带来的深刻影响和冲击,提出了整体性治理要求。从学理上来说,法律与伦理在规范价值层次、调整范围、规范方式和强制程度等方面存在很大差异,但又是相互联系、相互影响[12],二者的转化存在一定条件限制。首先,伦理准则所负载的价值有层次高低之分,并具有不同的适用效力,而法律在复杂关系调整和多元利益调适中必须保持“中立”、“公正”及“普适”,所以法律只可能吸收伦理规范体系中最基本的内容和要求[13]。在人工智能伦理治理领域,“隐私”、“数据安全”等准则所负载的价值属于低层次规范,作用边界清晰且适用效力高,可以确认为法律规范来实现规制,我国先后出台了《个人信息保护法》、《数据安全法》作为响应;而“科技向善”、“人民福祉”等伦理准则属于高层次的伦理价值,因其立法效应难保证、在实践应用中难抓取而容易被束之高阁,很难完成“软硬”转化。其次,伦理作为法律的基本来源与重要补充,在一般情况下,主要是指伦理的精神实质和价值取向为法律所选择或者吸收,一般都要对伦理规范进行降格化和具体化的立法技术处理[12],提升了转化难度。

4 结论与讨论

在新兴前沿技术进入发展快车道并广泛渗透的背景之下,科技伦理准则为科技伦理治理框架的搭建提供了顶层价值,为合理控制风险提供了重要指引,成为引导科技伦理治理技术生态、组织生态及制度生态构建的软体系,但准则的实践适用性却引起广泛争论。本文以人工智能伦理准则为切入口,选取全球范围内代表性伦理准则,利用文本计量的方法,瞄准准则发展的时间轴,完成了准则关键词的阶段对比差异分析、准则核心主题的聚合体系分析,进而总结了世界范围内人工智能治理共识的6大演变趋势,实践导向被凸显出来。

人工智能技术内生和应用外生的伦理风险具有明

显的特殊性，其负面效应无法立即显现且难以直接量化，风险来源较为复杂，智能治理经验普遍缺乏。伦理准则对AI技术研发及应用的行为指导和规约还未能对接，伦理准则与制度层面的立法、立规还不能有效衔接，伦理治理要求与企业主体可信AI实践之间还存在偏差。本文从组织、技术及制度三个层面分析了人工智能伦理准则中关键核心原则的关系，并进一步探讨了人工智能伦理准则的实践困境。

对于技术层面存在的“准则可操作性受限”及“准则考量要素不同质”的实践困境，其症结在于准则中所规定的规范性目标缺乏实际影响，因而要通过改进或补充准则以提升有效性。可以尝试从抽象的伦理价值和原则中推导出具体的技术实现内涵，列出问题清单，进而补充更详细的技术解释，将开发、实施和使用AI系统的实践与伦理所设定的价值观和原则关联起来。

对于组织层面存在的“伦理洗白”可能，要将伦理准则与人工智能企业的技术研发活动、产品开发及部署应用环节对应起来，可尝试在可信AI的关键实践中系统定义符合伦理的设计方法和框架，并通过组织制度和流程再造予以固化，以填补AI准则与其在应用实践之间的差距。要推进组织全员化的系统伦理教育，将伦理规范有效内化并融入企业文化。打破AI研发的封闭环境，人工智能企业要与公众务实沟通AI技术及产品可信的边界，重视用户同意、隐私及透明度。推动人工智能行业自律并提倡合作共建，实施政府优先采购导向，奖励并宣传典型实践案例，鼓励行业可信AI实践的发展。

对于制度层面存在的“准则与法律转化”困境，政产学研各界要深刻意识到，相较于具有社会指向、影响范围大且发挥“底线规定”功能的法律来说，伦理具有鲜明的“非强制性、非约束性”特征，会根据不同技术发展阶段、不同群体的认知波动而不断动态演进，发挥着“高线锚定”功能，主要依托各利益相关主体自愿、无约束力的合作行为且没有具体的执行机制，二者的转化需要一定的条件，只能做到相对同化，适度转化，可逐步建立相互补充的治理框架。

作者贡献声明：

贾婷：设计研究框架；文献梳理；数据整理分析；论文撰写；图表绘制；论文修订；

陈强：提出研究选题；明确总体研究目标；完成理论基础搭建及概念辨析；协调研究过程；

沈天添：理论模型推导；数据处理；数据分析；论文撰写。

参考文献：

[1] WU W, HUANG T, GONG K. Ethical principles and governance technology development of AI in China[J]. Engineering, 2020, 6(3): 302.

[2] MCNAMARA A , SMITH J , MURPHY H E . Does ACM′s code of ethics change ethical decision making in software development? [C]// the 2018 26th ACM Joint Meeting (ESEC/FSE). [S.L]: Association for Computing Machinery, 2018: 729 - 733.

[3] BENKLER Y. Don’t let industry write the rules for AI[J]. Nature, 2019, 569(7755): 161.

[4] 贾开，薛澜. 人工智能伦理问题与安全风险治理的全球比较与中国实践[J]. 公共管理评论，2021，3(1)：122.
JIA K, XUE L.Governance of ethical challenges and safety risks of artificial intelligence: global comparisons and practice in China [J]. China Public Administration Review, 2021, 3(1) : 122.

[5] 中国科学院自动化研究所.LAIP-链接人工智能准则平台[EB/OL].[2023-01-02]. https://www.linking-ai-principles.org/
Institute of Automation, Chinese Academy of Sciences. LAIP-linking artificial intelligence principles platform [EB/OL].[2023-01-02]. https://www.linking-ai-principles.org/.

[6] 陈小平. 人工智能伦理导引[M]. 合肥：中国科学技术大学出版社，2021.
CHEN X P. Ethical guidance on artificial intelligence[M]. Hefei: University of Science and Technology of China Press, 2021.

[7] 杨慧，杨建林. 融合LDA模型的政策文本量化分析——基于国际气候领域的实证[J]. 现代情报，2016，36(5)：71.
YANG H, YANG J L. Quantitative analysis of policy text merged with LDA model based on the field of international climate as demonstration[J]. Journal of Modern Information, 2016, 36(5): 71.

[8] VAKKURI V, KEMELL K K, KULTANEN J, *et al*. The current state of industrial practice in artificial intelligence ethics[J]. IEEE Software, 2020, 37(4): 50.

[9] KHAN A A, BADSHAH S, LIANG P, *et al*. Ethics of AI: a systematic literature review of principles and challenges[C]//The International Conference on Evaluation and Assessment in Software Engineering . Gothenburg Sweden: Association for Computing Machinery, 2022: 383-392.

[10] VAKKURI V, KEMELL K K, TOLVANEN J, *et al*. How do software companies deal with artificial intelligence ethics? a gap analysis[C]//The International Conference on Evaluation and Assessment in Software Engineering. Gothenburg Sweden: Association for Computing Machinery, 2022: 100-109.

[11] MITTELSTADT B. Principles alone cannot guarantee ethical AI [J]. Nature Machine Intelligence, 2019, 1(11): 501 .

[12] 刘华. 法律与伦理的关系新论[J]. 政治与法律，2002(3)：2.
LIU H. A new theory of the relationship between law and ethics [J]. Politics and Law, 2002(3): 2.

[13] 徐向华. 中国立法关系论[M]. 杭州：浙江人民出版社，1999.
XU X H. The theory of legislative relations in China [M]. Hangzhou: Zhejiang People′s Press, 1999.