人工智能伦理治理标准化指南

（2023 版）

国家人工智能标准化总体组
全国信标委人工智能分委会

二〇二三年三月

编写单位（排名不分先后）

中国电子技术标准化研究院
上海商汤智能科技有限公司
之江实验室
中国人民大学
中国科学院软件研究所
郑州中业科技股份有限公司
北京百度网讯科技有限公司
蚂蚁科技集团股份有限公司
天津大学
英特尔（中国）有限公司
北京理工大学
西南政法大学
山东人工智能研究院
第四范式（北京）技术有限公司
海信集团控股股份有限公司
云从科技集团股份有限公司
威麟 AIII 人工智能国际研究院
上海依图网络科技有限公司
广东中科凯泽信息科技有限公司
广州柏视医疗科技有限公司
中科软件测评（广州）有限公司
公安部第三研究所
北京眼神科技有限公司
马上消费金融股份有限公司
上海燧原科技有限公司
重庆中科汽车软件创新中心
深圳市北科瑞声科技股份有限公司
浙江大学
上海人工智能实验室
暨南大学
厦门大学
阿里巴巴（中国）有限公司
华东师范大学
华为技术有限公司
清华大学
中俄数字经济研究中心
中国医学科学院生物医学工程研究所
武汉大学
IBM
北京小米移动软件有限公司
北京旷视科技有限公司
深圳云天励飞技术股份有限公司
山东浪潮科学研究院有限公司
西安交通大学
国家计算机网络应急技术处理协调中心
中科南京软件技术研究院
OPPO 广东移动通信有限公司
特斯联科技集团有限公司
上海计算机软件技术开发中心
北京瑞莱智慧科技有限公司
深圳鲲云信息科技有限公司
国科础石（重庆）有限公司
深圳市矽赫科技有限公司
杭州量安科技有限公司

编写组成员（排名不分先后）

范科峰	董　建	徐　洋	杨雨泽	潘恩荣	蒋　慧	王迎春
杨嘉帆	古天龙	余　霄	郭　锐	李　介	徐　浩	胡正坤
孟令中	李斌斌	王祥丰	瞿晶晶	吕明杰	彭骏涛	金　博
王同益	杨　舟	包沉浮	柳嘉琪	王岚君	蒲江波	马　锐
王　骞	陈　亮	倪士光	程海旭	王海宁	李　锐	刘　佳
于　磊	魏简康凯	高雪松	李　军	袁　杰	翁家良	赵春昊
吴　军	高　卉	李　龙	王阳露	鲍　薇	马珊珊	孙　宁
贾一君	马骋昊	蒋哲琪	施锦诚	李慧杰	郭建领	李晓翠
晏　奇	卢紫梦	肖哲晖	张　琦	宋文林	叶　珩	沈　超
颜子夜	王　含	杨天元	刘继顺	陈文捷	张亚浩	张伟强
张一鸣	李云峰	栾丽红	李云峰	洪宝璇	赵芸伟	王思善
马朝阳	王小璞	涂小芳	方黎明	黄石磊		

目　　录

前言

世界科技发展历程中，重大技术变革往往带来生产力、生产关系及上层建筑的显著变化，为人类社会带来积极影响的同时，也带来对伦理的深刻反思。

目前，以深度学习为核心的新一代人工智能技术取得了极大的成功，大模型的发展令人工智能在下游任务的性能逐步趋近类人智能，并体现出极强的应用赋能潜力。人工智能新技术正在不断刷新着人们的认知极限，颠覆性地重塑着人类生活、工作和交流的方式，与人类社会融合为一。但是，人工智能产业保持高速发展态势的同时，人工智能技术自身发展面临诸多困境。人工智能所带来的隐私泄漏、偏见歧视、责权归属、技术滥用等伦理问题已引起“政产学研用”各界的广泛关注，人工智能伦理成为无法绕开的重要议题。因此，如何确保人工智能研发及应用符合人类伦理，让人工智能更好地造福社会、被公众信任是管理主体和研发主体等利益相关方必须积极解决的问题。

2022年3月20日，中共中央办公厅、国务院办公厅印发了《关于加强科技伦理治理的意见》(中办发〔2022〕19号)，明确提出了新时期科技伦理治理基本要求和重点任务。为进一步推动落实文件的相关要求，由中国电子技术标准化研究院牵头，依托国家人工智能标准化总体组和全国信标委人工智能分技术委员会组织浙江大学、上海商汤智能科技有限公司等政产学研用等50余家单位共同编制完成。

《人工智能伦理治理标准化指南》共分为六章，以人工智能伦理治理标准体系的建立和具体标准研制为目标，重点围绕人工智能伦理概念和范畴、人工智能伦理风险评估、人工智能伦理治理技术和工具、人工智能伦理治理标准体系建设、重点标准研制清单，以及展望与建议等6个方面展开研究，力争为落实人工智能伦理治理标准化工作奠定坚实基础。

编写组

2023年3月16日

1 概述

人类社会于20世纪中后期进入信息时代后，信息技术伦理逐渐引起了广泛关注和研究。信息技术的高速变革发展，21世纪后人类社会迅速迈向智能时代，随着人工智能的发展,越来越多的人工智能赋能应用、智能机器人等人工智能产品走入人类生活,人工智能可直接控制物理设备，亦可为个人决策、群体决策乃至国家决策提供辅助支撑；人工智能可以运用于智慧医疗、智慧工厂、智慧金融等众多场景，还可能被用于武器和军事之中。人工智能技术被日益广泛地应用在社会各个场景之中，甚至成为人类社会的一部分，研究人工智能伦理为系统反思人类既有伦理体系提供了重要契机。

现有人工智能技术路径依赖大量人类社会数据，特别是反映了人类社会演化历程中积累了系统性道德偏见的人类语言数据的训练，这样的人工智能系统进行的决策将不可避免地做出隐含着道德偏见的选择。然而，迈向智能时代的过程如此迅速，使得我们在传统的信息技术伦理秩序尚未建立完成的情况下，又迫切需要应对更加富有挑战性的人工智能伦理问题，积极构建智能社会的秩序。

技术与伦理正如两条相互缠绕的通道指引着人工智能的健康发展，一面展示着人类科技认知的水平,另一面展示着人类道德文明的程度。因此，如何结合技术手段和治理体系，合理地对人工智能伦理问题进行限制，也是人工智能领域最值得探讨的议题之一。

1.1 人工智能伦理概念

人工智能伦理与数据伦理、机器人伦理、信息技术伦理等应用伦理学分支具有密切的关系，有其发展的继承性和很多的相似之处。对于尚处弱人工智能的当前阶段，也很难判定人工智能伦理已经从机器人伦理等前沿

研究领域中实现完全分化。站在第四次工业革命的宏观角度来看，人工智能作为具有创造性和革命性的新领域，对全社会、全行业和全技术都在进行新的赋能，正在产生颠覆性、不可逆的后果，因此人工智能伦理也超越了应用伦理。因此，包括计算元伦理（Computational Meta-ethics）在内的创新学科，为人工智能伦理学理论体系的丰富做出不同理论进路的贡献。

随着人工智能的伦理问题逐渐引起人们的关注，越来越多的学者投入针对人工智能伦理的研究。基于IEEE Xplore的资源（截至2020年），国际社会对人工智能伦理诸多问题的关切主要集中在如下主题：人工智能的技术奇点问题、人工智能本身的伦理问题和人工智能对人类社会各领域造成的冲击与挑战从而带来的伦理问题。因此，可以总结出人工智能伦理内涵包含三方面：一是人类在开发和使用人工智能相关技术、产品及系统时的道德准则及行为规范；二是人工智能体本身所具有的符合伦理准则的道德编程或价值嵌入方法；三是人工智能体通过自我学习推理而形成的伦理规范。由于目前仍处弱人工智能时代，对于最后一点的讨论还为时尚早，从而，基于人工智能技术的伦理反思和基于伦理的人工智能技术批判共同构成了人工智能伦理的基本进路，也是人工智能伦理体系下的两大主要知识脉络。

（1）伦理与道德的关系

“伦理”是人类实现个体利益与社会整体利益协调过程中，形成的具有广泛共识的引导社会人际和谐和可持续发展的一系列公序良俗，诸如向善、公平、正义等，其内涵会根据研究主体的特性而改变；“道德”则表现为善恶对立的心理意识、原则规范和行为活动的总和。

（2）伦理与科技的关系

科学技术是人对客观物质世界的运动及其相互关系和规律的认识并运用于生产实践的产物，从一开始就内嵌着人类伦理道德的成分。21世纪以

来，科学技术呈现高度复杂、数字化、智能化、虚拟性和多系统综合的特征，技术后果的影响具有间接性和深远性。现代技术不仅将自然作为干预和改造甚至控制的对象，也把人变成了改造、增强和控制的对象，甚至出现了人机共生等现象。更进一步，人工智能作为一种前沿的科学技术，不仅实现了对人类体力劳动的替代，也在越来越多地代替人类的智力劳动。因此，人工智能伦理将对人类社会现有的伦理体系产生颠覆性影响。

（3）伦理、科技伦理、人工智能伦理的关系

如图1所示，科技伦理是更一般化的人工智能伦理，因此对人工智能的伦理思考需要回归到科技伦理的分析框架下，确定人工智能伦理反思的出发点和着眼点。当前人工智能技术在全球快速研发与应用，由于人工智能技术本身特有的不确定性问题，相比传统信息技术更加具有伦理反思的必要。另外，科技伦理的反思要在技术、人、社会、自然等综合的关联环境中，针对科技进步的条件、使用技术达到的目的，实现目的采用的手段和后果，进行评价与治理。

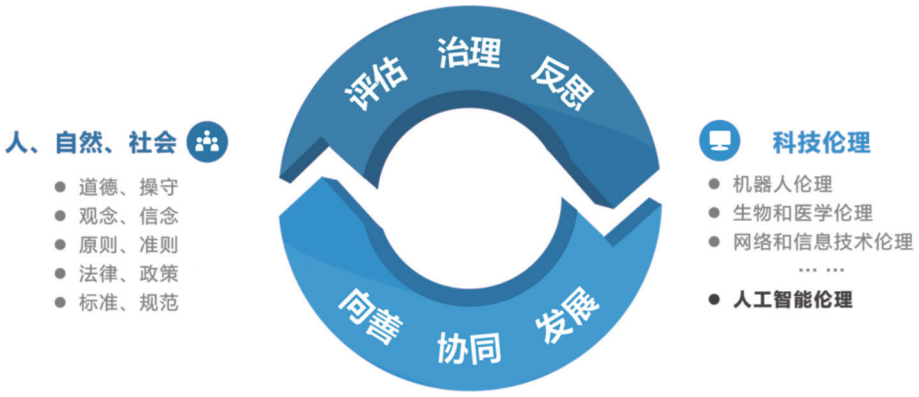


图1 人、自然、社会与科技伦理

1.2 人工智能伦理治理发展现状

世界各国和地区在大力推动人工智能技术突破和产业发展的同时，高度重视人工智能的全面健康发展，并将伦理治理纳入其人工智能战略，相应地推出政策或发布规划、指南与规范等文件用于建立人工智能伦理保障体系，开展人工智能伦理相关技术的管理。在激励人工智能产业发展的同时约束人工智能风险，体现了发展与治理并重的基本原则。

1.2.1 国际治理路线存差异，难以推动形成全球共识

a) **国际组织**

2019年5月，经济合作与发展组织（OECD）正式发布《人工智能原则》，该原则给出了人工智能可信赖的五项原则，包括可持续发展、以人为本、透明可解释、鲁棒性与信息安全、可问责等，并基于上述五项原则提出了政策建议和国际合作方针。

2021年11月25日，联合国教科文组织正式发布《人工智能伦理问题建议书》，该建议书作为全球首个针对人工智能伦理制定的规范框架，明确规定了4项价值观、10项伦理原则以及11项政策建议，其中人工智能伦理原则主要涉及相称性和不损害、安全和安保、公平和非歧视、可持续性、隐私权和数据保护、人类的监督和决定、透明度和可解释性、责任和问责、认识和素养、多利益攸关方与适应性治理和协作。

b) **欧盟**

欧盟认识到，加快发展人工智能技术与其积极推进的数字经济建设密不可分，而要确保数字经济建设长期健康稳定发展，不仅要在技术层面争取领先地位，也需要在规范层面尽早占据领先地位。

2019年4月，欧盟AI高级专家组（AI HLEG）正式发布《可信人工智能伦理指南》,在该指南中提出了实现可信人工智能的参考框架，在该框架中可信人工智能的基础由合法合规、伦理、鲁棒性三项相辅相成，必不

可少的要素构成。指南根据该三项要素提出了尊重人的自主权、无害化、公平性、可解释性等四项基本原则。此外，指南指出要实现可信赖的人工智能，必须将上述四个伦理原则转化为可实现的具体要求，用于参与人工智能系统生命周期的各个利益相关方，如开发人员、部署人员和最终用户，以及更广泛的社会层面。

2020年10月，欧盟委员会通过《人工智能、机器人和相关技术的伦理框架》决议，该框架的针对潜在高风险人工智能相关技术，该框架从第六条到第十六条等多个方面规范了伦理相关义务，其中主要包括以人为本，安全、透明、可问责，无偏见、无歧视，社会职责、性别平等，可持续发展，尊重个人隐私和补救权益七项原则。

2021年4月，欧盟委员会发布了立法提案——《欧洲议会和理事会关于制定人工智能统一规则（人工智能法）和修订某些欧盟立法的条例》，其主要内容包括对人工智能应用风险进行划分，将风险等级划分为不可接受风险、高风险、风险有限以及风险最低四个级别，以对人工智能系统进行分级管理，并明确监管部门和处罚要求，意图通过法律手段提高可信人工智能系统的发展。

c) 美国

美国基于国家安全的战略高度，强调人工智能伦理对军事、情报和国家竞争力的作用。以行政令为指导思想，从技术、标准、管理、应用等层面推动人工智能伦理的规范和使用。

2019年6月，美国国家科学技术委员会发布《国家人工智能研究与发展战略计划》以落实上述行政令，提出人工智能系统必须是值得信赖的，应当通过设计提高公平、透明度和问责制等举措，设计符合伦理道德的人工智能体系。

2019年10月，美国国防创新委员会（DIB）发布《人工智能准则：美国国防部（DoD）人工智能伦理使用推荐性规范》，该推荐性规范主要基

于现有美国宪法和战争法以及国际条款中的伦理参考框架，共提出负责任、公平、可追溯、可靠性以及可治理五项伦理原则。

d) **英国**

2016 年11月，英国科技办公室发布《人工智能:未来决策制定的机遇与影响》，报告中关注人工智能对个人隐私、就业以及政府决策可能带来的影响，并就处理人工智能带来的道德和法律风险提出了建议。

2018 年 4 月，英国议会下属的人工智能特别委员会发布《英国人工智能发展的计划、能力与志向》，提出了“人工智能不应用于削弱个人、家庭乃至社区的数据权利或隐私”的基本道德准则。

e) **德国**

2017年6月，德国联邦交通与数字基础设施部推出全球首套《自动驾驶伦理准则》，提出了自动驾驶汽车的20项道德伦理准则。特别针对无可避免的两难事故决策，规定不得存在任何基于年龄、性别、种族、身体属性或任何其他区别因素的歧视判断。

f) **日本**

2017年5月，日本人工智能学会发布了《日本人工智能学会伦理准则》，要求日本人工智能学会会员应当遵循并实践尊重隐私的原则。值得注意的是，该学会与纯粹的学术团体不同，除高校、科研机构和产业巨头外，还有科学技术振兴机构等政府部门参与。

g) **新加坡**

2020年1月，新加坡个人数据保护委员会发布《人工智能治理框架》，该框架提出了人工智能治理结构和方法，并针对人工智能伦理问题，总结梳理了具备一定共识的伦理原则，并将上述伦理原则融入人工智能治理框架中。

表1 人工智能伦理相关国际政策法规文件

国家地区	发布机构	文件题目	发布时间	关键内容
国际组织	经济合作与发展组织（OECD）	《人工智能原则》	2019年5月	人工智能可信赖
	联合国教科文组织	《人工智能伦理问题建议书》	2021年11月	人工智能伦理
美国	美国计算机协会公共政策委员会	《算法透明性和可问责性的声明》	2017年1月	可解释性
	美国参议院	《2018年恶意伪造禁令法案》	2018年12月	深度伪造技术
	美国白宫	《美国人工智能倡议》	2019年2月	对抗样本技术、深度伪造技术
	美国国家科学技术委员会	《国家人工智能研究与发展战略计划》	2019年6月	公平性、透明性、可问责
	美国国会	《2019年深度伪造报告法案》	2019年6月	深度伪造技术
	美国国防创新委员会（DIB）	《人工智能准则：美国国防部（DoD）人工智能伦理使用推荐性规范》	2019年10月	负责、公平性、可追踪、可靠、可控
	美国政府问责局	《人工智能：联邦机构和其他实体的问责框架》	2021年6月	可问责
	美国国会	《算法责任法案》	2022年2月	公平性、透明、可问责
欧盟	欧盟委员会	《通用数据保护条例》	2018年5月	可解释性、公平性
	欧盟委员会	《算法责任与透明治理框架》	2019年4月	透明、可问责
	欧盟AI高级专家组（AI HLEG）	《可信人工智能伦理指南》	2019年4月	人工智能可信赖
	欧洲议会研究中心	《人工智能伦理：问题和倡议》	2020年5月	人工智能伦理
	欧盟AI高级专家组（AI HLEG）	《可信AI评估列表》	2020年7月	人工智能可信赖
	欧盟委员会	《人工智能、机器人和相关技术的伦理框架》	2020年10月	人工智能伦理
	欧盟委员会	《人工智能法》	2021年4月	风险分级

国家地区	发布机构	文件题目	发布时间	关键内容
德国	德国联邦交通与数字基础设施部	《自动驾驶伦理准则》	2017年8月	自动驾驶伦理
英国	英国科技办公室	《人工智能:未来决策制定的机遇与影响》	2016年11月	隐私保护
	英国议会下属人工智能特别委员会	《英国人工智能发展的计划、能力与志向》	2018年4月	隐私保护
日本	日本人工智能学会	《日本人工智能学会伦理准则》	2017年5月	人工智能伦理
	日本内阁府	《以人类为中心的AI社会原则》	2018年12月	人工智能伦理
新加坡	新加坡个人数据保护委员会	《人工智能治理框架》	2020年1月	人工智能伦理
	新加坡个人数据保护委员会	《人工智能治理测试框架和工具包》	2022年5月	人工智能伦理

1.2.2 我国发展与治理并重，积极促成国际治理合力

我国将人工智能伦理规范作为促进人工智能发展的重要保证措施，不仅重视人工智能的社会伦理影响，而且通过制定伦理框架和伦理规范，以确保人工智能安全、可靠、可控。

2017年7月，国务院印发的《新一代人工智能发展规划》提出“分三步走”的战略目标，掀起了人工智能新热潮，并明确提出要“加强人工智能相关法律、伦理和社会问题研究，建立保障人工智能健康发展的法律法规和伦理道德框架”。

2019年6月，中国国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》，提出了人工智能治理的框架和行动指南。治理原则突出了发展负责任的人工智能这一主题，强调了和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理等八条原则。

2021年9月25日，国家新一代人工智能治理专业委员会发布《新一代

人工智能伦理规范》，该规范提出了增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养等6项基本伦理规范。同时，提出人工智能管理、研发、供应、使用等特定活动的18项具体伦理要求。

2022年3月20日，国务院办公厅印发《关于加强科技伦理治理的意见》，为进一步完善科技伦理体系，提升科技伦理治理能力，有效防控科技伦理风险，该意见提出应加强科技伦理的治理要求、明确科技伦理原则、健全科技伦理治理体制、加强科技伦理治理制度保障、强化科技伦理审查和监管以及深入开展科技伦理教育和宣传。

表2 人工智能伦理相关国内政策法规文件

发布机构	文件题目	发布时间	关键内容
国务院	《新一代人工智能发展规划》	2017 年 7 月	人工智能伦理
中华人民共和国工业和信息化部	《促进新一代人工智能产业发展三年行动计划(2018-2020 年)》	2017 年 12 月	AI 框架安全漏洞、AI 隐私风险
国家新一代人工智能治理专业委员会	《新一代人工智能治理原则——发展负责任的人工智能》	2019 年 6 月	人工智能伦理
全国人民代表大会常务委员会	《中华人民共和国个人信息保护法》	2021 年 8 月	可解释性
国家新一代人工智能治理专业委员会	《新一代人工智能伦理规范》	2021 年 9 月	人工智能伦理
国家互联网信息办公室	《互联网信息服务算法推荐管理规定》	2021 年 12 月	公平性、透明
中共中央办公厅、国务院办公厅	《关于加强科技伦理治理的意见》	2022 年 3 月	科技伦理
国家互联网信息办公室	《互联网信息服务深度合成管理规定》	2022 年 11 月	深度合成技术、透明
中华人民共和国外交部	《中国关于加强人工智能伦理治理的立场文件》	2022 年 11 月	人工智能伦理

2 人工智能伦理准则

2.1 概述

公众对科技发展持乐观态度且保持高位稳定状态，在充分肯定科技给人类带来好处的同时，越来越关注技术所带来的严重后果和社会风险。“加强科技伦理治理，实现高水平科技自立自强”的价值观是我国人工智能伦理准则的战略支撑之一。党的十九届五中全会提出了坚持创新在我国现代化建设全局中的核心地位，把科技自立自强作为国家发展的战略支撑。2022年3月20日，中共中央办公厅、国务院办公厅印发了《关于加强科技伦理治理的意见》，提出“科技伦理是开展科学研究、技术开发等科技活动需要遵循的价值理念和行为规范，是促进科技事业健康发展的重要保障。”，并明确了五大类科技伦理原则。

（一）增进人类福祉。科技活动应坚持以人民为中心的发展思想，有利于促进经济发展、社会进步、民生改善和生态环境保护，不断增强人民获得感、幸福感、安全感，促进人类社会和平发展和可持续发展。

（二）尊重生命权利。科技活动应最大限度避免对人的生命安全、身体健康、精神和心理健康造成伤害或潜在威胁，尊重人格尊严和个人隐私，保障科技活动参与者的知情权和选择权。使用实验动物应符合“减少、替代、优化”等要求。

（三）坚持公平公正。科技活动应尊重宗教信仰、文化传统等方面的差异，公平、公正、包容地对待不同社会群体，防止歧视和偏见。

（四）合理控制风险。科技活动应客观评估和审慎对待不确定性和技术应用的风险，力求规避、防范可能引发的风险，防止科技成果误用、滥用，避免危及社会安全、公共安全、生物安全和生态安全。

（五）保持公开透明。科技活动应鼓励利益相关方和社会公众合理参与，建立涉及重大、敏感伦理问题的科技活动披露机制。公布科技活动相

关信息时应提高透明度，做到客观真实。

本部分以《关于加强科技伦理治理的意见》中的五大类科技伦理原则为基础，总结归纳目前国内外人工智能伦理准则的关键词，梳理细化十类可实施性较强人工智能伦理准则，并基于标准化视角拆解、具象化各准则的要求，为后续具体标准的研制提供可操作的方向，具体细分准则及关键域见表3。

表3 科技伦理原则对应人工智能伦理准则

《关于加强科技伦理治理的意见》科技伦理原则	人工智能伦理准则	关键域
（一）增进人类福祉	（1）以人为本（For Human）	福祉、尊严、自主自由等
	（2）可持续性（Sustainability）	远期人工智能、环境友好、向善性等
（二）尊重生命权利	（3）合作（Collaboration）	跨文化交流、协作等
	（4）隐私（Privacy）	知情与被通知、个人数据权利、隐私保护设计等
（三）坚持公平公正	（5）公平（Fairness）	公正、平等、包容性、合理分配、无偏见与不歧视等
	（6）共享（Share）	数据传递、平等沟通等
（四）合理控制风险	（7）外部安全（Security）	网络安全、保密、风险控制、物理安全、主动防御等
	（8）内部安全（Safety）	可控性、鲁棒性、可靠性、冗余、稳定性等
（五）保持公开透明	（9）透明（Transparency）	可解释、可预测、定期披露和开源、可追溯等
	（10）可问责（Accountability）	责任、审查和监管等

2.2 人工智能伦理准则内涵

2.2.1 以人为本（for human）

在科技活动中，“以人为本”可理解为科技活动应坚持以人民为中心的发展思想。进一步聚焦到人工智能领域，“以人为本”的涵义包括：

第一，符合人类的价值观和伦理道德，尊重人权和人类根本利益诉求，遵守国家或地区伦理道德。应以保障社会安全、尊重人类权益为前提，避免误用，禁止滥用、恶用；

第二，遵循人类共同价值观，促进人机和谐，服务人类文明进步，促进人类社会稳健发展。坚持公共利益优先，推动经济、社会发展，不断增强人民获得感幸福感，共建人类命运共同体。

在不同应用场景中，“以人为本”涉及不同内容，可以归纳为福祉、尊严、自主自由三个次级关键词。

（1）福祉（well being）

“福祉”可能的应用场景包括：

（1）医疗健康。人工智能技术的设计者应确保满足对明确定义的使用案例或指示的安全性、准确性和有效性的监管要求；以健康需求为核心，提供实践中的质量控制措施并对使用人工智能改进质量提供有效测度。

（2）社会环境。根据公平性和预防损害原则，在整个人工智能系统的生命周期中，更广泛的社会、芸芸众生和环境也应被视为利益相关者，应鼓励人工智能系统的可持续性和生态责任。在理想情况下，人工智能系统应用于使所有人类（包括）后代受益。

（3）教育。教育人工智能的设计、开发和应用应当为教育活动的所有利益相关者带来福利，以实现所有教育教学活动相关人员的教育利益最大化为目标。

（2）尊严（dignity）

“尊严”可能的应用场景有：

（1）教育。人工智能的发展应维护儿童的尊严，重视并尊重儿童自身的思想、意愿、情感、兴趣爱好、自尊心等，避免对儿童的人格尊严造成伤害。

（2）居家养老。人工智能应在发挥辅助居家养老功能时，保障居家老人的自主意愿，维护其人格与尊严，避免加剧老年人自卑感、无力感与孤独感。

（3）智能机器人。人类尊严以社会关系为特征，要求用户明确是否以及何时与机器或另一个人交往，因此在智能机器人的使用过程中，必须保留将某些任务交给人类或机器的选择权利。

（3）自主自由（autonomy and freedom）

“自主自由”可能的应用场景包括：

（1）医疗健康。人类自身应确保继续掌控医疗决策过程和对医疗系统的控制，自行选择人工智能介入治疗的方式和程度。

（2）教育。教育人工智能在给学习者提供认知支架和智能化教学支持服务的同时，应当保留学习者根据自身需要和个性特征做出选择的权利，而不是代替学生做出选择，即给予学生充分的自主选择权，而不是将其作为某种目标对象来对待。

（3）司法。自由自主原则即“在用户控制下”的原则，要求除特殊规定外，必须确保用户是知情的行动者，并能够自主做出选择。

2.2.2 可持续（sustainability）

“可持续性”的基本涵义为一种可以长久维持的过程或状态。可持续性指人们在满足人类需求与未来发展时，在资源开发、技术发展和制度变革中保持环境平衡与和谐的过程，也指生态和社会关系中能够保持的一定

的过程或某种可接受的状态。在不同人工智能伦理准则中，“可持续性”的内涵包括：

第一，人工智能的设计和发展应该是为了促进社会和人类文明的进步，促进自然和社会的可持续性，使全人类和环境受益，提高社会和生态的福祉；

第二，人工智能系统的目标应该被清楚地确定和证明。理想情况下，人工智能系统应该被用来造福全人类，包括后代；

第三，帮助人工智能在全社会领域发展，让人工智能可以全方位立体性地渗透人们的生活，让每个人都可以享受人工智能带来的发展成果。

在不同应用场景中，“可持续性”涉及不同内容，包括远期人工智能、环境友好、向善性等。

（1）远期人工智能（long term AI）

未来十年是人工智能产业发展的重要时期，远期人工智能技术的发展路径将会沿着算法、算力两条主线向前发展，并逐步带领人类进入到人机协同时代。发展人工智能的最终目标不是要替代人类智能，而是要与人类智能形成互补，使得人类从繁重的重复性工作中解放出来，从而专注于推动人类自身文明进步。人工智能的发展关系到整个社会、全人类和环境的未来，因此在人工智能的研究、开发、使用、治理和长期规划中，应呼吁其健康发展，以支持建设一个拥有共同未来的人类社会，并实现有益于人类和自然的人工智能。

（2）环境友好（environment friendly）

人工智能可以从多个方面对环境产生积极影响，到2030年，人工智能在环境方面的应用能够产生约5.2万亿美元的价值。人工智能可以帮助减少人工成本，进行精准调控，有效降低生产风险，降低对环境的有害影响，实现可持续性的发展目标。

例如，人工智能技术可以将生活生产有关的设施集成，构建高效的设

施设备、人员、资源的管理系统，实现远程控制设备、设备间互联互通、设备自我学习等功能，通过收集、分析用户行为数据，为用户提供个性化服务的同时，综合管理水电等资源，实现环保节能的环境。

（3）向善性（prevention of harm）

在全生命周期中，人工智能系统的相应预测结果或行为应不超出实现合法目的或目标所需的范围，不得对人类、个别社区和整个社会造成损害。人工智能系统在数据、算法、应用管理三个层面均应考虑向善性的因素。

数据层面的向善性聚焦对于数据处理活动的管理和治理，降低由于数据安全以及质量问题导致算法决策失误、偏见和恶意，以及考虑解决数据泄露而导致的个人隐私问题。算法层面的向善性要求从设计开发阶段就应该考虑算法决策的是否会存在对人、社会层面的潜在危害，并在系统立项阶段就需要针对算法安全进行评估。应用层面的向善性更多侧重在算法用户是否科学、合理、适度的使用人工智能系统，从而降低由于错用、误用、滥用等使用问题导致的偏见歧视、人身伤害等问题的发生。

2.2.3 合作（collaboration）

“合作”的基本涵义为互相配合做某事或共同完成某项任务。在人工智能伦理领域，建立跨文化互信是全球和谐发展的基石，实现人工智能对全球有益需在人工智能伦理标准与治理的诸多相关领域达成国际合作。因此，人工智能应“重视开放协作，鼓励跨学科、跨领域、跨地区、跨国界的交流合作……在充分尊重各国人工智能治理原则和实践的前提下，推动形成具有广泛共识的国际人工智能治理框架和标准规范。”

（1）跨文化交流（cross-cultural communication）

来自不同文化背景或国家的团体协力合作，确保人工智能技术的发展、应用、治理能够造福社会。跨文化交流对于实现相关的伦理及治理举

措至关重要。跨文化研究合作与交流有助于化解合作障碍、增进不同观点和共同目标的理解与信任。具体事例包括但不限于：

●各国人工智能研究人员合作完成项目，采取安全可靠的方式发展人工智能系统；

●建立各类沟通渠道，确保着眼于人工智能伦理问题的国际讨论能够平等汲取多样的国际视角；

●邀请各国利益相关者参与制定实践准则、标准及法规等。

人工智能伦理与治理关乎全球人工智能发展与创新的方向与未来，跨文化交流重点在于厘清需要国际准则或协定进行规范的问题，或者确定需要突出文化差异的情况。为建立更牢固的跨文化信任，要正确认知、消除误解，增进不同文化与国家之间的互相理解，更好地推动相关原则、政策、标准、法律的制定、技术与社会落地。

（2）协作（cooperation）

人工智能的数字鸿沟问题指新技术和能力的掌握者在生产力、生活水平、文化教育等多个方面与其它群体显著拉开差距，严重情况下造成社会或国家间的两极分化。因此，人工智能技术先进国家应以当地语言开发人工智能伦理教育的在线课程和数字资源，并考虑到环境多样性，特别要确保采用残障人士可以使用的格式，促进人工智能技术技能教育与人工智能教育的人文、伦理和社会方面的交叉协作。应鼓励在人工智能领域开展国际合作与协作，以弥合地缘技术差距，填平“数字鸿沟”。应在充分尊重国际法的前提下，与其民众之间、公共和私营部门之间以及技术上最先进和最落后的国家之间，开展技术交流和磋商。

2.2.4 隐私（privacy）

“隐私”的基本涵义指的是自然人的私人生活安宁和不愿为他人知晓的私密空间、私密活动、私密信息。保护隐私具有维护人格尊严、维护个

人安宁、提高个人安全感和保护个人自由等作用。人工智能伦理准则中涉及隐私的条款包括：

●尊重隐私。人工智能发展应尊重和保护个人隐私，充分保障个人的知情权和选择权。在个人信息的收集、存储、处理、使用等各环节应设置边界，建立规范。完善个人数据授权撤销机制，反对任何窃取、篡改、泄露和其他非法收集利用个人信息的行为。

●保护隐私安全。充分尊重个人信息知情、同意等权利，依照合法、正当、必要和诚信原则处理个人信息，保障个人隐私与数据安全，不得损害个人合法数据权益，不得以窃取、篡改、泄露等方式非法收集利用个人信息，不得侵害个人隐私权。

在人工智能伦理中，“隐私”这一重要的关键词，还涵括以下内容：

（1）知情和被通知（informed and notified）

在全球个人信息保护制度中，“知情同意”原则一直是基础性制度，也是人机交互的核心问题。“通知和同意”是公平信息惯例（FIPs）的关键组成部分，是20世纪末为应对信息日益数字化而制定的一套原则。越来越多的国家或地区通过立法建立数据保护和隐私制度，核心是为合法收集和处理有关个人的数据提供法律依据，强调个人对信息的控制和数据处理的程序保障，而不是对做法的实质性禁止。

在线隐私和信息共享影响着人类生活的方方面面，“同意”概念和征求同意的机制反映出现有机制忽视伦理和规范价值问题。现有的“知情同意”或“通知同意”机制面临着保护人权的挑战。理想的解决方案既应最大程度地访问数据，又应保护每个人控制隐私和数据使用透明性的权利，保留所有人撤销同意访问的权力。

（2）个人数据权利（personal data rights）

个人数据权利包括获得数据和掌控数据的权利。获得数据的权利是指人们可以通过合法方式、途径或渠道，无障碍获取各种信息的能力，包括

主动获取信息与被动接收信息的利益；这一权利的实现，前提在于获取信息的方式、途径或渠道必须符合法律规定，任何偏离法律的信息获取，或是通过非法手段获得的信息，都不属于人们正常享有的获得信息的权利。同时，个人享有对于其个人数据的权利，并有权获得适当程度的保护，使得个人可以掌控其个人数据。个人对其个人数据的权利包括知情权、同意权（决定权）、查阅权、删除权、可携带权、更正权、补充权等。

（3）隐私保护设计（privacy protection design）

人工智能算法基于个人数据描绘用户画像，基于特定画像类型进行相关内容、产品或服务的匹配。人工智能获取个人数据并进行处理，从而做出与数据主体高度匹配的决策。因此，需要限制人工智能处理个人数据的能力，防止人工智能决策损害人的合法权益，通过对决策所需个人数据收集和处理的规制，或者是对算法公开化、透明化、可解释性等方式的规制，限制人工智能对个人数据不合理、不合法和不必要的处理。

确定隐私设计原则被认为是最好的选择，主要包括七大原则：积极预防，而非被动救济；隐私默认保护；将隐私嵌入设计之中，使其成为系统的核心组成部分同时又不损害系统的功能；功能完整——正合而非零和，主张实现用户、企业等多方共赢；全生命周期保护，主张为用户隐私提供从摇篮到坟墓全过程保护；可见性和透明性；尊重用户隐私，确保以用户为中心。

2.2.5 公平（fairness）

“公平”概念有着极为丰富的内涵，是社会建构、蓬勃发展的基础要素。从基本含义看，公平指处理事情合情合理，不偏袒哪一方。在社会生活中，认识和评价是否公平往往具有明显的主观色彩，人们容易从特定立场和目的出发，选择不同的标准和尺度进行评判。

人工智能活动至少应确保在平等主体的交往中各方充分交流、相互知

情，在共识机制下，分享人工智能带来的益处，合理分配风险；在社会组织中，为所有人提供同等、没有偏颇的机会，以及合理分配社会资源和利益；在集体、民族国家的交往中，确保各方共享人工智能技术红利，并尊重不同的文化与习俗。典型的公平问题包括大数据杀熟、算法黑箱、加剧认知茧房等对用户利益进行侵害的行为。

公平作为人工智能伦理准则的一个重要关键词，具体涵义同以下关键词密切关联：

（1）公正（justice）

“公正”指通过公平合理的利益（权利、权力、财富、机会等）分配使每一个成员得其所应得。公正是一种价值要求，要求分配的公平性与合理性。在人工智能领域，公正也往往涉及社会层面的资源分配，以及对社会不良现象的正义的惩罚，对受害者适当的保护。公正的可能应用场景包括：社会执法、司法裁判、政务服务、社会有限资源的合理分配。

（2）平等（equality）

在社会的组织过程中，存在诸多利益分配关系的对等。人工智能活动下的平等更多强调在社会资源、利益分配方面的机会均等，或不同主体间应当存在的对等关系。如保证社会各类主体获得教育、物资、工作、技术等与其发展息息相关的机会均等，保证不同背景的群体都可以按其习惯或舒适的方式使用公共的人工智能服务。

（3）包容性（inclusion）

“包容性”多针对技术应用阶段，指技术开发或服务提供者在给终端用户提供包含人工智能能力的服务时，需要确保人工智能技术的惠益人人可得可及，同时也要考虑不同年龄、不同文化体系、不同语言群体、残障人士、女童和妇女以及处境不利、边缘化和弱势群体或处境脆弱群体的具体需求。

（4）合理分配（reasonable distribution）

人工智能应用于社会时，会与不同形式的社会公平问题产生联系。人工智能应确保平等且正义地分配社会利益和损失，促进社会有限资源的合理分配，避免“数字鸿沟”，包括教育、医疗、应急、生活物资等；确保社会个体或群体不受到偏见、歧视、污名化，确保技术成果和带来的福利人人可享受。人工智能系统的使用不应让人们受到欺骗或不合理的损害，而应保障其自主决策权利。此外，公平也意味着人工智能从业者应遵守手段和目的的比例原则，审慎思考如何平衡利益和目的，即当有多种方法可以实现特定目的时，应考虑选择对基本人权和伦理规范产生负面影响最小的方案。

（5）无偏见与不歧视（unbias and non-discrimination）

人工智能系统本身或对其的应用可能因存在偏差或受到社会偏见影响，而引发或激化社会偏见，进而导致歧视行为。导致人工智能系统存在偏差的因素包括模型数据训练、模型结构设计、系统开发等。而导致系统应用存在偏差或偏见的因素较为复杂，多与使用者对系统的理解，以及其固有的社会偏见认知相关。

2.2.6 共享（share）

“共享”即分享，基本涵义是将一件物品或者信息的使用权或知情权与其他人共同拥有，有时也包括产权。在人工智能伦理领域，共享作为常见关键词之一，涵义包括：

- 包容共享，人工智能应促进协调发展，推动各行各业转型升级，缩小区域差距；应促进包容发展，加强人工智能教育及科普，提升弱势群体适应性，努力消除数字鸿沟；应促进共享发展，避免数据与平台垄断，鼓励开放有序竞争。
- 以自由、开放和共享为内核的共享伦理，其为一种超越物伦理的信

息伦理，即以自由、开放和共享为内核的共享伦理。

在人工智能的具体应用场景中，共享除基本含义外还应当包括平等。

（1）数据传递（data transmission）

随着对数据安全的重视和隐私保护法案的出台，曾经粗放式的数据共享受到挑战，各个数据拥有者重新回到数据孤岛的状态。同时，互联网公司更难以收集和利用用户的隐私数据，数据孤岛成为常态。实现数据的更充分利用，需要在满足隐私保护和数据安全的前提下，在不同组织、公司与用户之间进行数据共享。共享智能是希望在多方参与且各数据提供方与平台方互不信任的场景下，能够聚合多方信息进行分析和机器学习，并确保各参与方的隐私不被泄露，信息不被滥用。

鼓励算力平台、共性技术平台、行业训练数据集、仿真训练平台等人工智能基础设施资源开放共享，为人工智能企业开展场景创新提供算力、算法资源。鼓励地方通过共享开放、服务购买、创新券等方式，降低人工智能企业基础设施使用成本，提升人工智能场景创新的算力支撑。以物联网为基础，通过物联化、互联化、智能化的方式，综合无线传感技术、自动控制技术、网络技术和数据库技术实现现代化、智能化、共享化管理。以大数据智能为基础，需要解决数据碎片化的问题，实现从数据到知识，从知识到智能的跨跃，打穿数据孤岛，建立链接个人和机构的跨领域知识中心，形成开放式、互联互通的信息共享机制。为实现数据的有效共享，建议构建开放共享的领域大数据云平台。对各级机构、各种信息数据源的信息进行统一管理，实现对个体数据的高度整合。

另一方面，在现有信息化平台的基础上进行标准化改良，统一数据格式和描述规范，实现不同机构、不同来源信息存储与表达的规范化。利用标准化信息接口串联各机构数据，优化信息管理结构，实现信息系统的实时、同步更新，实现各级、各机构间的信息共享网络。

（2）平等沟通（equally discussion）

“平等”原则一般出现在人工智能招聘、选拔等对人的筛选活动中。人工智能招聘工具不仅可以帮助企业更加快速地对大量应聘者进行初步筛选，节省人力资源工作者的时间。然而，人工智能招聘工具在获得广泛信任之前面临的最大障碍之一是缺乏公共数据。一方面，机器学习的数据是非公开的，人们无法确认提高招聘中算法公平性的努力是否真的有效。另一方面，一旦使用人工智能工具歧视某些群体被证明，公司可能会面临严重的法律后果。另外，由于目前人工智能算法可解释性较低，当歧义出现时，无法合理向落选者给出原因。

人们在经济、政治、文化等方面处于同等的地位，享有相同的权利，考虑平等就业的重要性，因此，在招聘过程中，以及在可能影响劳动权利、消费者权利的场景下使用人工智能应用将始终被认定为“高风险”。需要给予人工智能算法受用者一个人与人之间平等沟通的机会，确保人的意志被充分传达。

2.2.7 外部安全（security）

“主动安全”的基本涵义是信息系统不受来自系统外部的自然灾害和人为的破坏，防止非法使用者对系统资源，特别是信息的非法使用的特性。在人工智能伦理中，安保在不同人工智能伦理准则中的具体涵义有：

1.在人工智能系统的整个生命周期内，应避免并解决、预防和消除意外伤害以及易受攻击的脆弱性，确保人类、环境和生态系统的安全。人工智能系统的主动安全准则主要侧重在人工智能系统的性能表现和抵抗外部恶意攻击的能力，该能力与人工智能可控性与可问责性相关，增强人工智能系统的可控性和可问责性可以有效提升系统的安全性。

2.增强安全透明。在算法设计、实现、应用等环节，提升透明性、可解释性、可理解性、可靠性、可控性，增强人工智能系统的韧性、自适应

性和抗干扰能力，逐步实现可验证、可审核、可监督、可追溯、可预测、可信赖。

“主动安全”在不同应用场景中还包括：

（1）网络安全（cyber security）

网络安全可能出现在对抗机器学习（adversarial machine learning）这一应用场景中，例如：数据投毒，即破坏训练数据完整性、保密性、可用性；或者是对抗样本攻击，即通过物理世界或数字世界对模型进行攻击，降低模型性能。对应的网络安全防范措施为对抗训练，可以有效削弱数据集中下毒攻击和对抗样本攻击的负面影响。

（2）保密（confidential）

在人工智能数据全生命周期中，数据采集、数据预处理（数据脱敏、数据标注）、模型训练、模型部署、模型推理等环节存在数据安全问题，可能导致个人信息泄露。因此，保密性在人工智能数据全生命周期中都值得关注。

（3）风险控制（risk control）

风险控制存在于人工智能模型研发全流程。在设计阶段需要考虑设计需求、风险管理、算法安全与数据安全，进行算法安全评估、个人信息安全影响评估以及风险评估。在人工智能模型研发全流程中，对算法、数据等内容需要从保密性、完整性、可用性、稳健性、隐私性等原则进行风险识别、风险分析，并进行风险评估，最后给出风险控制措施。

（4）物理安全（physical security）

物理安全包括人工智能系统相关硬件安全，以及涉及硬件层与相关应用使能层的安全模块，如通信加解密、存储加密、攻击检测、鉴权管理、密钥管理、硬件根密钥、可信执行环境、安全审计等。物理安全主要针对人工智能系统、计算设备等硬件。

（5）主动防御（active defense）

主动防御过程包括确认（validation）、验证（verification）和测试（test）三方面。确认主要针对机器的运行行为进行确认，以确保其符合用户的需求。验证针对机器的软硬件实现进行验证，以确保其符合预先定义的规范，同时这一规范又是人所定义的。测试针对机器的运行过程及结果进行测试，以保证其符合测试者预先定义的目标。

2.2.8 内部安全（safety）

安全的基本涵义是平安无损害，不受危险或损害的。根据 ISO/IEC TR 15443 1:2012《信息技术安全技术 信息技术安全保障框架 第一部分:总揽和框架》对安全进行定义，“安全”是指对某一系统，据以获得保密性、完整性、可用性、可核查性、真实性以及可靠性的性质。

在不同人工智能伦理准则中，“安全”一词被频繁提及，具体可以分为以下几点：

（1）可控性（controllability）

安全准则下的可控性，又可称为人类控制或技术的人为控制。最基本的要求是确保人工智能永远处于人类控制之下，在此基础上，要求个人（专家，员工，业务人员）的判断，融入人工智能系统流程中去，即人在环路（human-in-the-loop）。

（2）鲁棒性（robustness）

鲁棒性指的是在机器遇到故障或干扰的情况下，保持其功能在一定程度上的稳定性。在特定干扰或故障下可接受的鲁棒性的指标需要人为设定。鲁棒性应以准确性为前提，即机器可接受的准确性指标需要人为设定。

（3）可靠性（reliability）

可靠性主要指机器需要与其预设目标一致地运行，当遇到严重问题

时，机器自身或者机器与人的交互需要确保具有退路（fallback）方案。可靠性还包括韧性（resilience），即人工智能系统应具备一定的抵抗恶意攻击脆弱性的能力，包括数据投毒和模型参数窃取等攻击方式。

（4）冗余（redundancy）

在人工智能系统全业务场景的应用中，冗余包括设备冗余、系统冗余、体系冗余。所谓冗余，即从安全角度考虑额外的数量，如通过多重备份来增加系统的可靠性，或者在传输信息时，借助于信息的重复和信息的累加使对方能够得到明确的信息，以避免遭受信道（channel）和噪声的干扰。只有在人到机的全体系都做到冗余，才能尽量降低安全风险。

（5）稳定性（stability）

人工智能系统的稳定性，主要包括四个方面：第一，状态一致性，即输入输出的关联性；第二，人的稳定性，如体系、团队的稳定；第三，机器自身的稳定性，如软硬件、系统、代码的稳定；第四，人与机器交互的稳定性，如接口稳定、人机交互的稳定。

2.2.9 透明（transparency）

在人工智能伦理领域，人工智能的透明性是指在不伤害人工智能算法所有者利益的情况下，公开其人工智能系统中使用的源代码和数据，避免“技术黑箱”。透明度要求在因知识产权等问题而不能完全公开算法代码的情况下，应当适当公开算法的操作规则、创建、验证过程，或者适当公开算法过程、后续实现、验证目标的适当记录。透明度的目的是为相关对象提供适当的信息，以便他们理解和增进信任。具体到人工智能系统，透明度可以帮助人们了解人工智能系统各个阶段是如何按照该系统的具体环境和敏感度设定的。透明度还包括深入了解可以影响特定预测或决定的因素，以及了解是否具备适当的保证。透明原则在以下三种应用场景中尤其值得关注：（1）无法解释的算法；（2）训练的数据集透明度不足；

（3）训练数据选择方法的透明度不足。在不同应用场景中，透明原则具有不同内容：

（1）可解释（explainable）

端到端的深度学习，一个广为诟病的问题是其不透明性或不可解释性。随着神经网络模型越来越复杂，在准确性越来越高的同时，我们不得不在模型的准确性和可解释性之间做出妥协，两者常难以兼顾。结果是，有时候研究人员做出一个行之有效的模型，却并不能完全理解其中的缘由。因此，即便得到的结论是正确的，也将因为不知道结论的推导过程而失去其作用。

（2）可预测（predictable）

“可预测性”，一方面要求确保人工智能系统是准确、可靠且可被重复试验的，提升人工智能系统决策的准确率，完善评估机制，及时减少系统错误预测带来的意外风险。另一方面要求严格保护人工智能系统，防止漏洞、黑客恶意攻击；开发和测试中最大限度地减少意外的、不可预测的后果和错误，在系统出现问题时可执行后备计划。

（3）定期披露和开源（reveal and open-source regularly）

披露和开源机制是人工智能治理的重要组成部分。它通过在被规制者与规制受益群体之间建立激励机制，以推动人工智能算法开发者和运用者增强自我规制的能力，从而在市场上获得更大的份额。政府监管部门可以要求人工智能算法企业和平台定期发布社会责任报告，强制披露风险活动，公开数据处理的目的、数据处理者的身份、数据接收者等相关信息。在当下发展中，算法越来越复杂，决策的影响却越来越重大,应积极鼓励人工智能开源活动。

（4）可追溯（traceable）

可追溯即确保人工智能决策的数据集、过程和结果的可追溯性，保证

人工智能的决策结果可被人类理解和追踪。可追溯性是支撑人工智能自身透明度、可责性、可审计性的关键质量特性，因此在设计开发与验证过程中，应考虑数据集、算法设计、算法需求规范、源代码、风险管理、算法测试六大要素及其内在关联，在运行时考虑使用环境、数据输入、决策过程、输出结论、使用反馈的可追溯，以过程运行日志文档、可视化界面和可追溯性矩阵的形式呈现。

2.2.10 可问责（accountability）

在人工智能领域，“责任”指的是人工智能系统应该履行对利益相关者的责任和社会责任，并且可追责。责任原则的具体建议包括采取诚信行为和厘清责任的归属两方面。

诚信行为指作为人工智能研发者在进行人工智能技术研发设计时应致力于创建高度可靠的人工智能系统，减少安全风险，不留“后门”；人工智能使用者在使用人工智能技术时应具有高度的社会责任感和自律意识，不利用技术侵犯他人的合法权益；人工智能技术相关各方应严格遵守法律法规、伦理道德和标准规范。人工智能技术不能成为限制责任的理由，必须明确人工智能情景下的责任，并根据因果关系进行规制。若有人使用人工智能技术侵害他人合法权利则需要承担相应责任，监管部门应将人工智能纳入各环节监管，将登记、审批、检查等方面落实到位。

根据不同应用场景，可问责性的具体涵义有所侧重：

（1）责任（responsibility）

责任存在于金融、教育、医疗、健康、体育、文化等众多场景中，指个体分内应做的事，来自职业要求、道德规范和法律法规等，当行业人工智能没有做好“分内”工作时，则应承担的不利后果或强制性措施。人工智能系统应该履行对利益相关者的责任和社会责任，并且可追责，不断评估和调参，以促进人工智能系统持续不断改进。责任还包括人工智能从业

人员和使用人员的认识和素养要求，例如了解人工智能系统的影响和风险等。

（2）审查和监管（investigation and supervision）

在人工智能使用及问责过程中，重视监督机构的审查和监管作用。应当建立企业内部或者外部的监督机构，明确人工智能系统相关方法定责任和义务，评估人工智能对人类权利的影响，以及对于技术负面影响的评估。同时保证人工智能在相同条件或者前提下，行为应该有一致性，并对相应的人工智能标准和最佳实践进行不断监督审视，确保个人可以对人工智能的决策进行申诉、纠错机制，强调对人工智能的自动决策提供补救措施。

3 人工智能伦理风险分析

人工智能技术作为使能技术，具有“头雁”效应，能够赋能千行百业，影响面广且影响程度深远。人工智能伦理风险也因此或凸显，或隐藏在各个人工智能技术被广泛应用的场景中。人工智能伦理风险可能导致各类伦理问题的出现，极端情况下可能引发重大公共安全问题等伦理危机。有必要对常见人工智能技术在具体场景中应用产生的伦理风险进行分类、识别与分析，以在人工智能技术从研发到工程转化和场景应用的全生命周期中提前规避和最大程度消解其伦理风险。

基于人工智能的敏捷治理原则，为促进研发机构和企业提供市场公众和政府监管等利益相关方理解、易用、敢用并用好的人工智能技术，本章节将对人工智能全生命周期中主要阶段的伦理风险进行分解分析，提供一个人工智能伦理风险分析和分类方法，并选取典型场景对其中的代表性伦理风险进行分析。

3.1 人工智能伦理风险来源

（1）数据

随着数据采集、机器学习、人工智能等技术的使用，数据集大小呈指数级扩大，数据富含越来越大的价值，从而也导致个人信息泄露的情况频繁发生。个人隐私保护、个人敏感信息识别的重要性日益凸现。人工智能技术需要海量数据用于训练算法，带来了数据盗用、信息泄漏等人工智能伦理风险。

在人工智能系统开发过程中，对数据集的代表性差、规模不清晰、均衡性不足等设计问题，易导致数据集的公平性受到影响，从而影响算法公平性；数据标注过程的数据泄露等数据预处理安全问题也会导致个人信息保护的问题；缺乏对数据层面的可追溯技术也会增加人工智能系统在责任

认定以及问责风险。

表4 数据层面人工智能伦理风险分析

开发环节	伦理风险分析	伦理风险属性
设计开发	数据主体采集授权相关风险	隐私保护
	数据集的规模、均衡性等设计不足	公平性
	数据预处理、模型训练等环节数据处理流程安全问题	隐私保护、安全
	数据预处理的质量问题，如数据标注准确率不足	公平性
验证测试	模型评估等环节数据处理流程安全问题	隐私保护、安全
	测试数据集规模、均衡性等质量问题	公平性
	测试数据集与训练数据集重复度高	透明及可解释性
部署运行	模型部署环节数据处理流程安全问题	隐私保护、安全
	模型部署时的数据集完整性等质量问题	安全
	模型部署时的数据集均衡性不足等质量问题	公平性
	模型推理环节运行数据处理流程安全问题	隐私保护、安全
	模型推理环节运行数据泄露、缺乏对运行数据的有效可追溯技术	可问责性
维护升级	再训练阶段数据处理流程安全问题	隐私保护、安全
退役下线	数据退役阶段数据泄露、留存数据未删除	隐私保护

（2）算法

在人工智能算法层面，主要存在以下3种形式的问题：存在因模型参数泄露或被恶意修改、容错率与韧性不足造成的算法安全风险；因采用复杂神经网络的算法导致决策不透明与无法被充分解释，同时数据的输入和输出关系理解不清晰，可能造成的可解释性安全风险；以及因算法推理结果的不可预见性与人类自身的认知能力受限，导致无法预测智能系统做出的决策原因与产生的效果，造成的算法决策偏见。

此外，算法在运行推理环节可能会被错误使用或滥用，如智能推荐算法一旦被不法分子利用，将使虚假信息、涉黄涉恐、违规言论等不良信息传播更加具有针对性和隐蔽性，在扩大负面影响的同时减少被举报的可能。

表5 算法层面人工智能伦理风险分析

开发环节	风险分析	伦理风险属性
设计开发	算法存在对特定人群以及性别歧视设计	公平性
	算法存在对个人和社会的恶意设计、潜在危害	向善性、可持续发展
	算法缺乏安全、可控性设计	安全、监督和决策
	模型训练环节的算法产生偏见、不公平问题	公平性
	模型训练环节的算法不可解释问题	透明及可解释性
验证测试	模型评估环节的算法偏见问题	公平性
	模型评估环节算法安全问题	安全
	模型评估环节中缺乏有效的版本管理、不可追溯等问题	可问责性
	模型评估环节的算法不可解释问题	透明及可解释性
部署运行	模型部署环节的环境不可控	监督和决策
	模型部署环节的算法安全、韧性问题	安全
	模型部署环节的算法不可解释问题	透明及可解释性
	模型推理环节的算法安全、韧性问题	安全
	模型推理环节的算法不可控问题	监督和决策
	模型推理环节的算法滥用、误用问题	监督和决策
	模型推理环节中缺乏有效版本管理、不可追溯等问题	可问责性
	模型推理环节的算法不可解释、不可预测问题	透明及可解释性
维护升级	模型更新时模型参数与配置不正确	安全
	模型更新时缺乏有效版本管理	可问责性
退役下线	模型退役时模型未彻底删除或模型参数泄露	隐私保护、安全

（3）系统（决策方式）

当前阶段，人工智能系统既承继了以往信息技术的伦理问题，又受到数据和算法风险的影响。由于人工智能在社会生产生活的各个环节日益广泛应用，人工智能系统固有的不透明性、低可解释性等特征，再加上系统漏洞、设计缺陷等风险，可能引发个人信息等数据泄露、工业生产线停止等社会问题，威胁个人权益、社会秩序、国家安全等。因此，人工智能系统可能在诸多方面带来一系列伦理风险，主要包括以下方面：人工智能系

统的缺陷和价值设定问题可能带来公民生命权、健康权的威胁，人工智能系统的滥用也可能威胁人身安全以及个人信息隐私权。比如，人工智能武器的滥用可能在世界范围内加剧不平等，威胁人类生命与世界和平；人工智能在工作场景中的滥用可能影响劳动者权益，并且人工智能对劳动者的替代可能引发大规模结构性失业的危机，带来劳动权或就业机会方面的风险。

（4）人为因素

人为因素主要体现在人为造成的算法歧视，主要分为两种：一种是由算法设计者造成的算法歧视，另一种是由用户造成的算法歧视。

算法设计者造成的算法歧视，是指算法设计者为了获得某些利益，或者为了表达自己的一些主观观点而设计存在歧视性的算法。这是因为算法的设计目的、数据运用、结果表征等都是开发者、设计者的主观价值与偏好选择，算法设计者是否能够将既有法律法规或者道德规范编写进程序指令中本身就值得怀疑。而设计开发者可能会把自己持有的偏见与喜好嵌入或固化到智能算法之中。这会使人工智能算法通过学习把这种歧视或倾向进一步放大或者强化，从而产生算法设计者想要的并带有歧视性的结果，最终导致基于算法的决策带有偏见。

由用户造成的算法歧视，主要产生于需要从与用户互动的过程中进行学习的算法，由于用户自身与算法的交互方式，而使算法的执行结果产生了偏见。这是因为在运行过程中，当设计算法向周围环境学习时，它不能决定要保留或者丢弃哪些数据、判断数据对错，而只能使用用户提供的数据。无论这些数据是好是坏，它都只能依据此基础做出判断。

3.2 人工智能伦理风险分析方法

人工智能伦理风险来自人工智能技术经过工程转化并进入具体场景形成人工智能技术应用后对人工智能伦理准则的冲击,人工智能伦理风险责任主体根据国家《新一代人工智能伦理规范》具体分为技术主体（对应研发活动）、应用主体（对应供应活动和使用活动），以及管理主体。结合本指南第三章节对人工智能伦理准则的分析，人工智能伦理风险分类识别与分析矩阵如图2所示。人工智能伦理风险具体分为应用型、技术型和（技术应用）混合型三类人工智能伦理风险，分别面向不同的风险责任主体。人工智能伦理风险主要治理原则对应第三章节所列的十大人工智能伦理准则,对号代表各伦理风险类别所需要主要考虑的治理准则，其中：

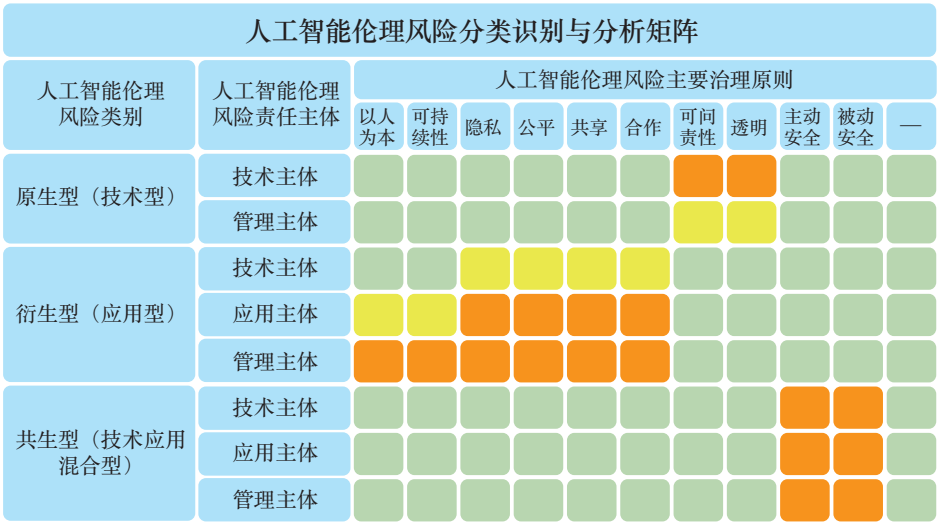


图2 人工智能伦理风险分析矩阵

（1）应用型人工智能伦理风险定义为：在人工智能技术应用过程中，由于人工智技术的工具放大效应，使得应用场景中的原有伦理共识中已经接受的伦理问题产生放大或者变形。这类问题并不是人工智能技术本

身的问题，而是技术应用带来的衍生效应。例如：智能教育过程中人工智能学习工具的使用，将人类社会原有的“数字鸿沟”推向“智能鸿沟”，进一步扩大了教育公平问题甚至加剧了社会的不平等。对于衍生型伦理风险，主要从以人为本、可持续性、隐私、共享、合作和公平这六个伦理准则的维度进行具体分析。应用型伦理风险的责任主体主要是应用主体和管理主体，分别指从事人工智能产品与服务相关的生产、运营、销售、采购、消费、操作等供应活动和使用活动的自然人、法人和其他相关机构等，和从事指人工智能相关的战略规划、政策法规和技术标准制定实施，资源配置以及监督审查等管理活动的自然人、法人和其他相关机构等。对此，应用主体应在供应方面尊重市场规则、保障用户权益，在使用发面避免滥用误用、积极提高使用能力；管理主体应推动敏捷治理、积极实践示范、正确行权用权、促进包容开放。

（2）技术型人工智能伦理风险定义为：由于人工智能技术自身路线与发展特点的伦理治理问题，由此带来的伦理风险通常与人工智能的技术特征直接相关，是人工智能技术应用给人类社会治理拓展的新的伦理边界。例如：深度学习技术应用于智能驾驶领域的环境感知和行为决策，现阶段其感知和决策结果的不可解释性导致驾驶者和乘客难以安心使用，也影响着其他交通参与者的安全，比如辅助驾驶系统突如其来的莫名刹车。对人工智能技术应用的信任程度及其原理解释问题成为新的人机物的关系问题，是典型的原生型伦理风险。对于原生型伦理风险，主要从可问责性和透明两个伦理准则的维度进行具体分析。技术型伦理风险的责任主体主要是技术主体，指的是从事人工智能相关的科学研究、技术开发、产品研制等研发活动的自然人、法人和其他相关机构等。对此，技术主体应加强技术研发和治理，逐步实现对人工智能技术的可验证、可审核、可监督、可追溯、可预测、可信赖。

（3）（技术应用）混合型人工智能伦理风险定义为：由于人工智能

技术在场景中的应用，使得安全这一固有的伦理共识被技术本身的特征与技术的场景应用共同影响，改变了人类对安全这一伦理价值的认知。即人工智能技术应用一方面增加了对安全程度的需求，另一方面也为安全新增了需要考虑的方面，最终形成技术型与应用型伦理风险交互、混合的情况。例如：知识图谱技术相关应用在提高了数据安全程度要求的同时，也新增了防范知识投毒方面的安全维度要求。对于混合型伦理风险，主要从主动安全和被动安全两个伦理准则的维度进行具体分析，其责任主体包括技术主体、应用主体和管理主体。对此，技术主体应积极增强人工智能系统的韧性和抗干扰能力；应用主体在供应方面应加强质量管控，在使用方面禁止违规恶用；管理主体应加强风险防范。

对人工智能伦理风险进行分类识别的目的是，有助于各责任主体和利益相关方清晰认识人工智能伦理风险产生的原因，并有助于各方进一步理解人工智能伦理风险与传统伦理问题的区别和联系，明确各责任主体的责任。进而，在人工智能伦理风险对伦理准则产生影响的维度上进行具体分析，有助于指导和促进人工智能技术和技术应用合乎伦理的研发，并有助于各方用好人工智能技术。本章节后续部分将选取典型场景，对其中的代表性人工智能技术应用产生的伦理风险进行分析。同时，附录1借鉴ISO/IEC TR 24368，为各方提供了一份构建和使用符合伦理设计的人工智能的非详尽考虑事项清单，可作为组织管理过程或信息系统审计中的部分参考，旨在减轻人工智能技术应用全生命周期中的伦理风险。

3.3 人工智能技术应用和典型场景伦理风险分析

本小节将选取自动驾驶、智能媒体、智能医疗、只能电商、智能教育和科学智能等6个典型场景，使用上述人工智能伦理风险分类识别和分析矩阵，对场景中人工智能技术应用的伦理风险进行分析，在人工智能技术和技术应用的管理、研发、供应和使用中作为积极的构成要素，有助于在

人工智能技术应用的伦理风险发生前预判、规避和化解。

3.3.1 自动驾驶

自动驾驶即指车辆在搭载先进传感器、控制器、执行器的基础上，在特定的设计运行范围内，能自主获取和分析车内外信息，持续地处理部分或全部动态驾驶任务，包括单车智能和车路云深度协同等。自动驾驶是人工智能技术最受社会公众关注的应用场景之一。

第一类技术型伦理风险的代表，是环境感知、行为决策和运动控制等自动驾驶应用对可问责性这一伦理准则的影响。自动驾驶的终极目标是自主决策实现无人驾驶，但由于道路、行人和天气情况的复杂多变，计算机视觉算法的不够成熟可靠，以及目前传感设备识别能力有限等问题，导致现阶段自动驾驶汽车事故频发。事故原因多元而难以追溯，并且现有的制度规范也未建立完善，为自动驾驶新增了问责困难的伦理风险。

第二类应用型伦理风险的代表，首先是自动驾驶汽车的操控权和自主决策行为对以人为本、合作和公平等伦理准则的影响。自动驾驶系统对车主操控权的适时归还以及车主的适当使用等问题已经引发了伦理争议。面对“电车难题”等不可避免的道德困境，自动驾驶系统代替人类做出的选择，可能隐含着数据、算法、系统和人为歧视，并可能把个体的偏差放大到社会面。此外，由于不同国家和地区的文化背景和价值观不同，自动驾驶系统很难有一个普适的道德责任框架，这对国际合作和统一提出了挑战。其次是驾驶数据包括导航数据对隐私这一伦理准则产生的影响，主要是座舱数据对驾驶员和乘员的人脸、声纹、指纹等生物识别特征数据的采集。车外数据对交通参与者人脸和车牌等个人信息的采集，以及导航数据对行驶轨迹的采集，也侵犯了其他交通参与者的隐私。

第三类混合型伦理风险的代表，是驾驶数据、车联网和车路协同对主动安全和被动安全两大伦理准则的而影响。海量驾驶数据的价值已然巨大，而重要敏感区域的地理信息、车辆流量、汽车充电网等能够反映经济

运行情况的数据更关系到国家安全和公共利益，对安全提出了更高要求。此外，车联网和车路云协同使得黑客能够远程操控、影响和攻击车辆，影响道路交通安全，并衍生高科技犯罪和破坏公共安全的伦理风险。

3.3.2 智能媒体

智能媒体场景关注人工智能技术在社交媒体方面的应用，能够提高信息传播效率、人机交互体验、内容质量与丰富程度。人工智能技术对社交媒体的赋能也带来了伦理风险。

第一类技术型伦理风险的代表，是智能内容处理和生成应用对可问责性和透明两大伦理准则的影响。深度合成技术特别是深度伪造的发展带来了虚假信息甚至虚假账号的可能性，使得海量的音视频和图像需要接受检测。因此，这些深度合成的内容通过各类社交媒体传播，导致信息本身和信息发布者的失真与难以分辨。进而，这些虚假信息可能衍生其他领域的伦理风险，比如影响公安领域的网络舆情监控有效性、导致特定人员声誉受损等。

第二类应用型伦理风险的代表，是智能网络社交应用对以人为本、公平和隐私等伦理准则的影响。网络社交的兴起使得原本只在家人朋友和同事邻里之间分享的私密生活，可以被用户线上展示分享并收获关注。但这首先带来了隐私泄露的伦理风险。其次，社交平台的推荐算法应用可能驱动这些内容甚至隐私向更多人开放，缩短了被他人获取和分析的链条，可能加剧造谣、污蔑、歧视甚至网络暴力等伦理风险，并衍生如高科技犯罪、诈骗和破坏社会信任关系等伦理风险。最后，社交平台的推荐算法机制可能使得伤害性和反智性言论更易扩散，并加剧“信息茧房”、“回音室”以及群体极化等伦理风险。

第三类混合型伦理风险的代表，是海量社交数据以及敏感信息对主动安全和被动安全两大伦理准则的影响。一方面，海量社交数据的产生对数据和网络安全提出更高要求。另一方面，一些政治或经济相关虚假信息的

传播，信息推送中特定词语插入对用户情绪和行为的干预与诱导，以及机密信息的泄露和扩散，也可能造成意识形态和社会舆论方面的伦理风险，甚至威胁国家安全。

3.3.3 智能医疗

智能医疗场景是人工智能技术在医学场景的赋能与应用。医学自身已是一个伦理敏感的场景，对于人工智能技术应用于医学，极有可能放大或变形传统医学场景的伦理风险、新增或产生混合的伦理风险，需要特别关注。

第一类技术型伦理风险的代表，是智能手术机器人和智能医学影像分析等涉及自主能动性的人工智能应用对可问责性这一伦理准则的影响。在智能手术机器人协助手术过程中因为医生操控失误或机器故障引起的医疗事故，现阶段仍没有对如何划分责任形成共识。智能医学影像分析等“AI+医疗”技术亦存在类似的责任划分困境，其分析结果对医生的正确判断可能产生的影响必须纳入考虑。

第二类应用型伦理风险的代表，首先是脑机接口应用对以人为本这一伦理准则的影响。脑机接口是人工智能和医学的代表性交叉技术，由于直接作用于人体，可能造成对大脑组织的创伤和感染，其安全和伦理风险和争议较大。另一方面，未来可能发生的黑客攻击和意念控制将导致人类自由意志受到巨大威胁，更是将有关人脑的手术对自由意志的可能影响的伦理风险进行了极度放大或变形。其次是可穿戴式设备、智能医疗信息平台等对隐私、共享和公平等伦理准则的影响。随着现代人对日常健康监测重视，这些设备或平台大量采集人体的生物特征数据并进行分析处理、辅助诊断和医学研究。使用者难以保障数据的删除权、存储权、使用权和知情权，且由于这些数据多是生理信息而显得更为敏感，将对患者隐私的侵犯可能性进行了放大。另外，智能医疗设备的高昂费用和使用门槛，也放大了原本的数字鸿沟和医疗资源不均衡等问题，对共享和公平两大伦理准则

产生进一步影响。

第三类混合型伦理风险的代表，是人工智能驱动的人类增强应用对主动安全和被动安全两大伦理准则的影响，脑机接口技术也在此列。可能的黑客攻击和干扰、数据窃取等对相关设备联网的主动安全提出了要求，而相关设备对人脑或人体的直接接触和影响也对其可控性、可靠性和鲁棒性等被动安全提出了要求。

3.3.4 智能电商

智能平台场景主要是人工智能技术赋能的电商和服务平台，能够为消费者在线提供各类产品和服务，做到个性化、精准化和高效化。

第一类技术型伦理风险的代表，是推荐系统对可问责性这一伦理准则的影响。由于用户一般只关注推荐排名靠前的商品，不良商家能够通过推荐和排序算法背后的一系列操作获得优先推荐的机会，可能导致电商平台陷入伦理争议并形象败坏。同时，平台和商家的责任划分，甚至消费者如何选择，均引发伦理争议。

第二类应用型伦理风险的代表，是推荐系统和智能调度系统等人工智能技术应用对以人为本、共享、公平和隐私等伦理准则的影响。一方面，电商平台使用的算法可能过度诱导消费并导致“信息茧房”，用户被反复加强和固化的消费偏好也可能破坏市场的有序竞争和创新活力。另一方面，平台可能根据用户的历史交易记录、交易习惯等特征，甚至“窃听”或“监视”聊天记录，利用算法在价格等方面实施不合理的差别待遇或进行针对性推送，造成不公平竞争甚至欺诈行为。如近年来被广泛曝光的“大数据杀熟”现象，便严重违反了消费者的合法权益，破坏了市场秩序。而服务平台则主要依靠智能调度系统为消费者提供各类服务如外卖、跑腿和用车等，虽然提高了效率，但缺失了人性关怀的温度，可能导致在算法控制与归训下的员工陷入为按时完成任务而采取违规违法行为的伦理困境，对自身、平台和公众均产生风险隐患。

第三类混合型伦理风险的代表，是消费数据和管制物品流通对主动安全和被动安全两大伦理准则的影响。一方面，海量且全面消费数据能够反映国家经济发展运行态势；另一方面，多样的电商平台伴随着现代物流服务，使得枪支、入侵生物、危化品等管制物品的流通难度和成本降低，且提高了监管和追缴难度。这些均对国土安全、经济安全等提出了主动安全和被动安全方面的更高要求。

3.3.5 智能教育

智能教育是人工智能技术在教育领域应用产生的新形态，有望对教育技术、教育目标、教育理念以及教育治理等产生变革性影响。教育是社会发展的动力源泉，智能教育场景的伦理风险需要进一步重视。

第一类技术型伦理风险在智能教育场景尚未凸显，现阶段主要体现在算法黑箱对透明这一伦理准则的影响。

第二类应用型伦理风险的代表，首先是智能化学习和评价应用对以人为本和公平等伦理准则的影响。学习者的独立思考和教育者的评价、关怀，对学习者的学习成效和积极性非常重要。但智能学习内容推荐应用可能导致学习者遭遇“信息茧房”而使知识结构和认知的片面化；“拍照搜题”可能惰化学习者思维和独立思考能力，违背教育教学规律等问题。智能教育缺失了教育的人文关怀，也可能因数据、算法或人为歧视而放大原本的人工偏差。其次是各类智能教育产品和服务特别是学习监测相关应用对隐私这一伦理准则的影响。学习者一方面可能意识不到数据和隐私被滥用或泄露，另一方面类似电子标牌、点阵笔和摄像头等用于监测学生上课和写作业情况的设备，更对学生形成了“环形监狱”，模糊家校界限并侵犯个人隐私。还有智能教育产品和服务对共享和公平两大伦理准则的影响。教育公平是最大的机会公平，智能教育产品和服务的费用和使用门槛，也放大了原本的数字鸿沟和教育资源不均衡等问题，对共享和公平两大伦理准则产生进一步影响。

第三类混合型伦理风险的代表，是智能教育技术对数据、算力和网络的需求而产生的对主动安全和被动安全两大伦理准则的影响。人工智能技术对教育领域的赋能大幅提高了教育数据的价值，也令更多教育资源暴露在网络空间中，从而对数据和网络安全提出了更高要求。

3.3.6 科学智能（AI for Science）

AI for Science场景指的是人工智能技术可有效赋能传统科学领域的研究，比如计算育种、计算材料学、计算生物学、计算天文学和计算社会学等等，同时包括了对人工智能技术自身的研究。

第一类技术型伦理风险的代表，是人工智能算法训练过程对透明这一伦理准则的影响。目前主流的人工智能算法一般不具有全局可解释性或仅具有局部解释性，被称作“算法黑箱”。深度神经网络是黑箱问题的代表性算法。“算法黑箱”直接催生了人们对透明的伦理需求，并进一步在具体应用场景对可问责性和透明等伦理准则产生影响。

第二类应用型伦理风险的代表，首先是人工智能算法训练过程对共享和公平两大伦理准则的影响。某些人为设计的价值取向或训练数据集的缺陷可能会引发整体的偏见和歧视问题。并且，由于这些偏见和歧视是隐藏在算法决策的黑箱中，人工智能算法可能会在不易被人察觉的情况下侵犯公众的正当权益、放大人类社会固有的偏见和歧视。其次是人工智能算法训练中的数据对隐私这一伦理准则的影响。人工智能模型的训练依赖大量的数据，然而，现阶段用于科研的数据获取、使用和存储尚未有明确的规范。一些科研机构用于科学研究的图像和数据，存在侵犯个人的肖像权、隐私权和知情权等的伦理风险。同时，这些敏感数据的存储与二次开发也带来了一定的伦理风险，可能被不法分子滥用。人工智能技术增加了暴力获取数据的可能性或降低了数据获取的成本，加剧了传统的大数据技术本身对隐私这一伦理准则的冲击。最后，现阶段人工智能技术对算力的需求也对可持续性这一伦理准则产生重要影响，最直观的伦理风险可能发生在

人工智能技术运用到环境治理时。一方面，人工智能技术赋能传统环保行业，助力解决可持续发展所面临的污染排放估算、资源合理分配等问题；另一方面，人工智能模型的训练、推理等计算也消耗着大量资源。更进一步的，为了提高人工智能模型的计算精度以及鲁棒性、泛用性等模型表现，“唯性能论”更倾向使用大规模预训练模型和规模更大的数据集，需要非常可观的计算资源，不可避免地大幅度增加能耗。

第三类混合型伦理风险在人工智能算法的训练过程中并未凸显，将在更为具体的场景中进行分析。

4 人工智能伦理治理的技术解决方案

新一代人工智能具有高度的自主性、自学习及适应能力等特征，给技术治理以及政府监管带来了新的挑战。人工智能有可能彻底改变世界，为社会、组织和个人带来诸多好处。然而，人工智能也可能会带来巨大的风险和不确定性。数据的不平衡和算法的局限性等技术性缺陷，都可能使人工智能受到偏差的影响，从而导致严重的伦理问题。因此，从技术上推进伦理准则实践落地是重要路径。伦理原则需要融入到人工智能技术的全生命周期中，例如在模型和算法在设计之初即需纳入相关伦理因素的考量，在训练过程中需要增强伦理原则的技术模块嵌入。

本章主要人工智能伦理的技术解决方案出发，总结当前主要的伦理落地的技术实践，从人工智能伦理技术框架、技术实现路径以及相关治理实践进行阐述。

4.1 人工智能伦理技术框架

人工智能伦理技术框架基于生命全周期可划分为4个阶段（见图3），其中阶段一与阶段二主要为人工智能产品上市前的研发阶段。阶段一包含人工智能产品的概念与设计两部分内容，人工智能伦理作为重要内容融入到概念和设计过程。阶段二为数据集和模型开发部分，通过基准数据集及相关诊断、人工智能伦理符合设计技术确保设计开发过程将正确的伦理观植入人工智能技术应用，以及诊断结果符合伦理规范。阶段三为上市前的产品市场准入阶段，包含了产品方的内部评测和外部评测两方面，通过人工智能评估评测技术从多个伦理维度进行分析。阶段四为应用推广和上市后的跟踪评估评测，通过应用反馈实现产品的优化。值得注意的是，以上四个阶段之间相互关联，存在多个反馈回路，伦理技术方案嵌入到人工智能产品全生命周期。

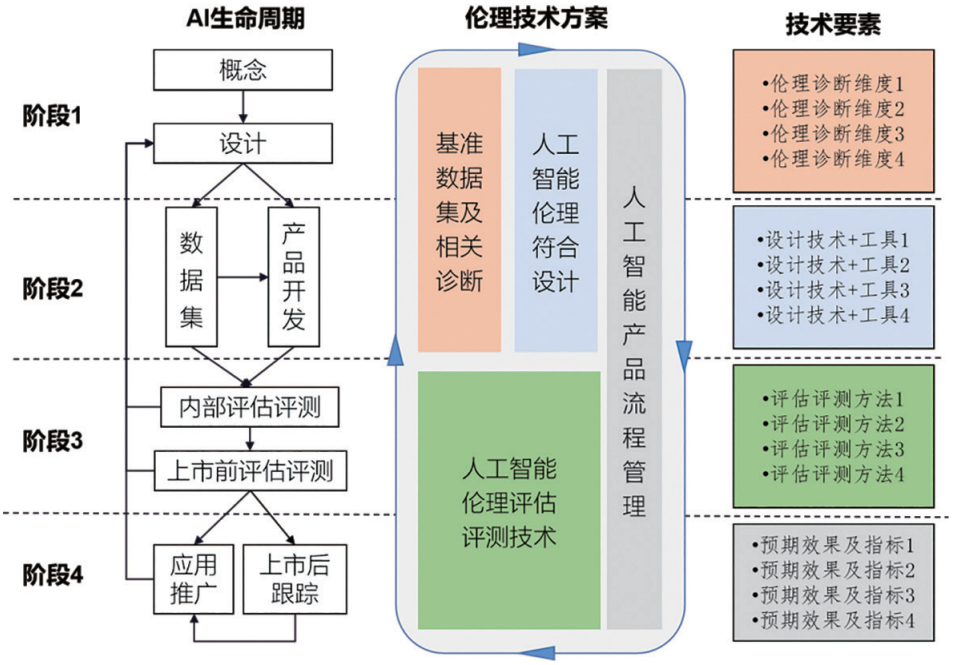


图3 人工智能伦理技术框架

相关基准数据集及诊断指标的合理性，从人工智能产品设计之初就已确定。目前相关的基准数据集以性别偏见、刻板印象偏见等为主，诊断指标包括公平性、隐私性和可解释性等。伦理嵌入技术是指将价值设计、隐私、公平、可持续等符合伦理准则的技术框架与路径采用嵌入方式加入技术模块。伦理评估评测（监管）技术是指在人工智能产品和应用在开发和推广过程中基于伦理原则进行技术审查。

从伦理技术效用角度，如公平性、鲁棒性、可解释性、隐私性可采用不同的技术和工具，应用于模型开发、内部评估评测、上市前评估评测、上市后评估评测等人工智能生命周期的不同阶段，使其满足相应指标要求（见表6）。

表6 人工智能伦理技术全生命周期赋能示例

伦理维度	模型开发→内部评估评测→上市前评估评测→上市后评估评测		
公平性	公平学习	公平性检测	
安全性与鲁棒性	对抗防御	鲁棒性检测	对抗检测
可解释性	可解释学习		可解释评估
隐私性	隐私学习	隐私攻击与检测	

目前，无论是伦理符合设计技术或是监管技术，均处于起步阶段，将伦理技术系统性融入人工智能创新还没有成熟的方法论。鉴于此，持续加大对人工智能伦理技术的研究探索，确保人工智能发展可信可持续具有重大意义。以下将重点聚焦分析隐私、安全性、透明与可解释性、公平典型伦理原则上的技术实现路径与治理实践。

4.2 人工智能伦理技术实现路径与治理实践

4.2.1 人工智能伦理技术实现路径

4.2.1.1 隐私性（Privacy）

数据隐私是人工智能伦理原则中最基本、最常见的要求。欧盟《通用数据保护条例》《非个人数据自由流动条例》《欧盟网络安全法案》等监管政策，以及国内《网络安全法》《数据安全法》均强调了数据隐私的重要性。人工智能作为基于数据驱动的科学研究范式，对基准数据集提供更高要求，强调数据集本身的准确性完整性和隐私性。

安全计算、联邦学习、同态加密等相关技术可以在不直接访问用户数据隐私的条件下，分布式进行人工智能模型的训练。例如，联邦学习不仅为机构间分布式机器学习模型的协同开发提供了一种用于隐私保护的技术解决方案，而且也为人工智能社区的可持续发展指明了一种新的商业模型，推动数字生态系统的可信任。此外，在预测阶段以及训练阶段的同态加密隐私保护技术、差分隐私保护技术，在保护机器学习以及深度学习中的用户敏感数据已经取得较大进展。

进一步地，在隐私攻击测试方面，针对成员推理攻击，允许访问模型输出的条件下，推断某样本是否属于训练集；针对属性推理攻击，在给定样本标签和部分特征的情况下，推断样本的其他特征，从而进一步确保人工智能模型训练的科学性和准确性。

4.2.1.2 **安全和鲁棒性**（security and robust）

随着人工智能算法和应用场景的复杂性增加，如何以经济高效且无差错的方式设计和实施一个安全和可信赖的系统，成为当前人工智能发展面临的一个巨大挑战。在安全与鲁棒性方面，数据投毒、后门攻击检测、伪装恶意样本、生成虚假样本影响数据集等现实问题层出不穷。为应对以上挑战，目前已形成对抗测试、博弈模型、形式化验证等多种解决方案。

（1）对抗测试。在系统研发全生命周期中加入对抗测试已经成为增强模型安全性与鲁棒性的主流方向。一般来说，对抗算法分为4个等级：随机攻击、盲盒攻击、黑盒攻击、白盒攻击。

（2）博弈模型。从博弈论的角度，将机器学习模型的交互过程建立为一个博弈模型，目标是找出一个均衡博弈状态（最优解），让防御者赢得博弈，从而提高机器学习模型的鲁棒性。以对抗样本生成和防御为核心的对抗深度学习是目前的研究热点。

（3）形式化验证。因为输入扰动地选择组合情况庞大，对抗测试无法列举出给定一组输入的所有可能输出，因此引入形式化验证作为对抗测试方法的补充十分关键。形式化验证包括完整验证程序和不完整验证程序：前者损耗较高成本来保证没有误报，但其扩展性有限；相比而言，后者泛化性较好，但是准确性较低。

4.2.1.3 **透明与可解释性**（Transparency and Explainable）

与机器学习中“黑匣子”概念相比，可解释的人工智能是一套流程和方法，可使人类用户能够理解和信任机器学习算法所产生的结果和输出。可解释的人工智能用于描述模型、其预期影响和潜在偏见，并且有助于描

述人工智能支持的决策中的模型准确性、公平性、透明度和结果。

可解释性主要包括数据可解释、特征可解释、模型可解释、逻辑可解释等方面。（1）数据可解释。通常称为深度模型解释，主要是基于数据分析和可视化技术，实现深度模型可视化，直观展示得到模型结果的关键依据。（2）特征可解释。指评估特征对模型的重要程度。（3）模型可解释。这类方法也称为可解释模型方法，主要是通过构建可解释的模型，使得模型本身具有可解释性，在输出结果的同时也输出得到该结果的原因，帮助人工智能工程师直观地打开模型“黑箱”。（4）逻辑可解释。目前的人工智能能力直接去“学习”人的逻辑难度很大，因此当前更多是尝试如何在建模型过程中融入人工经验，从而使得模型的产出与专家判断更吻合，比如端到端地对模型决策进行解释。

可解释性诊断指标主要包括保真度（Fidelity）、模型的复杂度、解释方法的类别、解释方法的可靠性、解释结果的正确性等方面。（1）保真度（Fidelity）。在事后可解释性中，符合黑盒模型的能力保真度是帮助利益相关者评估解释的一个基本属性。保真度的价值可以反映解释的有用性，而高保真的解释是一种有价值的解释方法的必要条件。（2）模型的复杂度。如决策树的深度、深度神经网络的层数等，对同一类模型而言，模型复杂度越高，可解释性越差。（3）解释方法的类别。类别主要包括建模前、建模中、建模后，一般来说建模中的可解释性最强，建模后次之，建模前最差。（4）解释方法的可靠性。解释方法需要经过人工智能专家的认证，确保其算法的可靠性。（5）解释结果的正确性。人工智能模型都有具体的应用场景，解释结果需要经过相关领域专家的认证，确保解释结果的正确性。

4.2.1.4 公平性（Fairness）

目前，人工智能公平性研究主要集中在评估不同群体之间或个人之间人工智能输出的差异。（1）个体公平。认为如果两个人有相似属性，则

人工智能算法应当做出相似决策。（2）群体公平。群体公平要求人工智能算法针对特定属性区分的用户群体要做出相同的概率预测，包括人口结构均等、概率均等和机会均等。由于此类公平不假定训练数据具有任何特殊属性，容易被验证。（3）反事实公平。在许多决策场景中，受保护的属性（如种族和性别群体）可能对预测结果产生因果性影响。

目前在人工智能公平性方面主要涉及公平性测试数据集和公平性机器学习设计。（1）公平性测试数据集及诊断指标。公平性测试数据集和普通数据集的差别在于具有敏感属性，目前国际上对公平性机器学习算法的测试大多基于美国司法部数据集、CrowS-Pairs（衡量刻板印象偏见）、Winogender（衡量与职业相关的性别偏见）、StereoSet（衡量性别、种族、宗教和职业上的刻板印象偏见以及原始语言建模能力的基准）等典型数据集。机器学习公平性诊断指标主要包括混淆矩阵、几率平等性（Equalized odds）、人口均等（Demographic parity）、不同误判率(disparate mistreatment, DM)。（2）公平性机器学习设计。从算法的基本定义出发，是在输入、过程及输出不同阶段描述解决问题的策略机制，包括反分类（anti-classfication)、分类均等（classification parity)、校准（calibration)。

此外，在公平性问题上，数据训练集的历史偏差、标注偏差，算法中的因果偏差、归纳偏差、属性偏差等，可以通过加强学习、差分学习、公平机器学习技术等技术措施以及例如社会学、统计学、法学、伦理学等多领域学科知识支撑进行偏误纠正。图4展示了公平性机器学习流水线。通常可以在三个阶段对模型进行去偏。第一，公平性预处理：在此阶段通常是对数据集进行校正，利用改造后的数据集进行训练；第二，公平性机器学习，主要是在模型优化过程中加入与公平性相关的正则项或者约束，使得训练出的模型无偏；第三，公平性后处理，主要是针对结果进行修正，主要用于输入数据和训练过程是黑盒的场景。

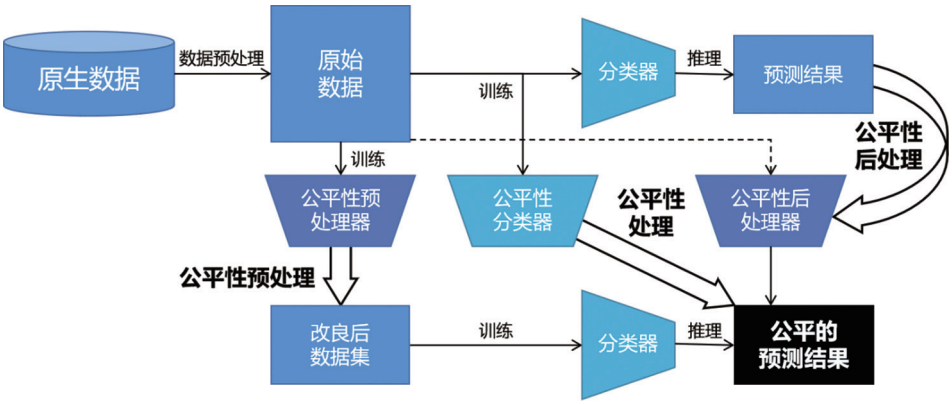


图4 公平机器学习流程

4.2.2 人工智能伦理管理实现路径

人工智能伦理原则融入技术研发和产品开发的全过程，不仅需要具备技术工具的保障，还应通过建立完善的人工智能伦理风险管理体系，确保伦理原则的实施和技术工具的使用贯穿于组织运行的全流程之中。伦理风险管理工具应服务于伦理风险管理流程中不同阶段的治理目标和治理要求，成为伦理治理理念和原则的有形载体。人工智能伦理风险管理工具的开发应紧密围绕组织的伦理原则开展，按照功能性，可分为以下几种主要类型：

4.2.2.1 伦理风险评估工具

“伦理风险评估工具”主要运用于人工智能系统的设计与开发阶段，主要包括伦理风险分级管理机制和伦理风险评估模板，详细信息如下：

（1）伦理风险评估模板

伦理风险评估模板应当基于统一的伦理风险评估框架，通过考察系统设计使用的场景、涉及的相关主体、预期实现的功能以及对社会和个人的影响，并结合场景、主体和功能定义人工智能系统全生命周期的风险点位和干预措施。

（2）伦理风险分级管理机制

综合欧盟《人工智能法案》（草案）以及美国国家标准技术研究院《人工智能风险框架》（草案）、加拿大《自动化决策指令》等相关人工智能风险分级思路，以及我国人工智能伦理相关政策指导文件，并结合人工智能产品开发和应用的实际情况，可总结出人工智能系统伦理风险分级参考原则，详见表7。

根据上述公开、明确的伦理风险等级，并结合个人权益、公平性、透明度、安全性等影响程度，建立伦理风险分级管理机制，帮助系统开发团队建立伦理风险清单；

表7 人工智能系统伦理风险分级表

伦理风险等级	伦理风险等级简介
E4	即禁止类系统，指背离人工智能伦理原则、违反法律法规要求的人工智能系统
E3	即伦理高风险系统，指直接关系最终产品安全、个人权益、市场公平、公共安全和生态安全的人工智能系统
E2	即伦理中风险系统，指对最终产品安全、个人权益、市场公平、公共安全和生态安全具有间接或潜在重要影响的人工智能系统
E1	即伦理低风险系统，指对最终产品安全、个人权益、市场公平、公共安全和生态安全不具备明显影响的人工智能系统
E0	即伦理无风险系统，不包含机器学习算法、不具备人工智能功能的人工智能系统

4.2.2.2 隐私性

“隐私保护管理工具”的使用贯穿于系统的全生命周期，其基本内容包括：

（1）个人信息安全影响评估：针对产品研发全生命周期流程，依据《中华人民共和国个人信息保护法》、GB/T 35273-2020等要求，对产品在进行数据收集、数据传输、数据存储、数据使用以及数据加工等数据处理活动进行个人信息保护自评估，明确产品在数据处理活动中应具备的功能，包括但不限于人工信息的无痕模式、去标识化处理、申请删除以及全

链路加密等功能；

（2）数据出境评估：针对存在数据出境的产品，依据《数据出境安全评估办法》等要求，对数据出境活动中双方的资质、传输目的和渠道、数据的规模、范围、种类、敏感程度进行评估。

4.2.2.3 公平性

“公平性管理工具”的使用贯穿人工智能系统的全生命周期，其基本内容包括：

（1）数据集公平性说明：用于说明所选取的数据集的完整性、可用性、具备充分的代表性等，从而降低由于数据集的缺陷、规模不足或者存在脏数据等情况所导致后续模型训练环节训练出来的模型存在偏见；

（2）系统运行机制说明：在系统开发过程中，应提供系统的运行机制说明文件，帮助用户理解系统用途、解释系统决策及可能存在的偏差；

（3）产品适用性说明：用于说明产品设计及相关功能在满足不同群体方面的考虑，是否有考虑弱势群体或十四岁及以下未成年人群体的 使用需求。

4.2.2.4 问责性

“问责性管理工具”的使用贯穿于系统的全生命周期，其基本内容包括：

（1）系统开发日志：系统开发的全流程应保持完整记录，并能够明确具体责任方；

（2）系统运营日志：系统上线应完整记录系统的操作、运行及客户使用和反馈信息。

4.2.2.5 透明与可解释性

“可解释性管理工具”的使用贯穿人工智能系统的全生命周期，并可结合披露对象的不同调整信息披露的形式和内容，其基本内容包括：

（1）披露对象识别要求：基于系统的伦理风险等级，以及相关法规

和政策要求，明确系统信息应披露的范围及要求；

（2）算法可解释性说明：用于说明所选取的算法类型是否具备充分的可解释性等，保障在开发设计阶段对算法的决策机制有一定的解释性说明文件，从而为算法的可解释性提供恰当、合理的说明；

（3）数据处理日志：系统开发过程中，为保证数据的可追溯性、完整性、可用性，可对数据处理活动进行记录，形成审计日志，包括但不限于数据采集、数据预处理、特征工程、模型训练、模型部署等过程；

（4）透明性功能检查清单：结合系统信息披露要求（如显著标识、更新提示等），设置上线前的功能检查列表。

4.2.2.6 安全性和鲁棒性

“安全性和鲁棒性管理工具”的使用贯穿于系统的全生命周期，同时应考虑与技术工具配合使用进行协同治理，其基本内容包括：

（1）算法分级备案管理：在系统开发过程中，应根据有关部门规定对算法进行分级备案管理；

（2）算法安全评估：并在设计开发阶段对算法进行安全评估，从人身安全、社会伦理、国家安全等方面评估算法的技术合理性和伦理安全；

（3）算法违法违规处置机制：针对算法违法违规事件，应设立算法安全应急管理机制和违法违规处理条例；

（4）数据安全管理机制：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》等数据安全法律法规和标准文件，设立数据安全管理机制，对人工智能系统开发过程中所涉及到数据处理活动进行规范。

4.2.3 人工智能伦理技术治理实践

人工智能伦理技术实践正在快速发展，国际组织、政府、企业等均探索相应的解决方案。从企业视角来看，全球典型的（互联网）平台型企

业积极推动人工智能伦理在产品和方案上落地应用。从政府和国际机构视角，包括经济合作与发展组织（OECD）、欧盟、中国、新加坡、德国在内相关主体在人工智能伦理技术实践上具有积极举措。

4.2.3.1 企业人工智能伦理技术治理实践

从人工智能全生命周期视角，评估评测主要包括上市前企业内部的评估，以及上市后的政府组织评估。在产品上市前，企业会对产品进行一系列内部测试与验证。目前，IBM、微软、谷歌、百度、华为等面向偏见检测、隐私保护、安全性、风险管理、提高公平性、透明度、可解释性等均开发了一系列内部测评工具开展自我审查工作与机制（工具清单见附表1）。例如， Google发布了名为What-If的交互式可视化工具，该工具能够以直观的方式检查负责的机器学习模型，以评估和诊断机器学习模型的公平性。此外，IBM、Facebook以及微软均发布了关于人工智能解释的工具。例如IBM 推出AI Explainability 360 工具包，集成了8种人工智能解释方法和2种评估指标。微软基于 Explainable Boosting Machine 算法开发的InterpretML，可用于训练可解释机器学习模型和黑盒模型，其使预测结果更精准，且有可解释性。

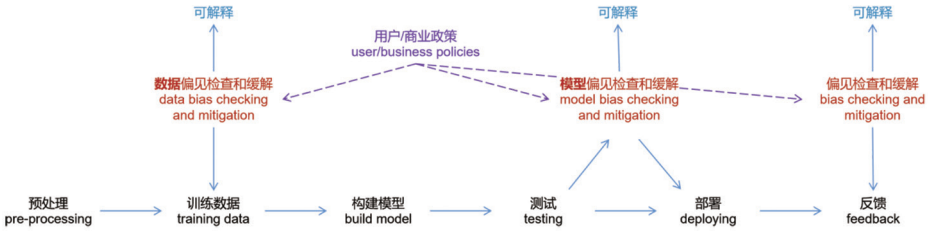


图6 产品伦理验证工具流程（pipeline）

4.2.3.2 政府和国际组织人工智能伦理技术治理实践

政府主要通过设立标准化评测评估平台、评估框架或者采用沙盒测试进行人工智能技术的监管。目前，由于人工智能伦理监管的敏感性，目前大部分政府和国际组织还处于观望状态。国际范围内没有出现专用的人工智能伦理测试工具或监管机制，人工智能伦理的考量一般作为人工智能系统应用整体评估的一部分，以透明度、公平性、偏见缓解等指标的形式体现。

2021年，OECD提出落地负责任人工智能原则的工具框架，从技术工具、过程性工具、教育工具对工具属性进行分类，然后给出了用于对比、描述具体工具的框架，该框架包含7个特征维度。该框架的目标，是希望帮助全球人工智能利益相关者更好的选择适合自己的工具，评估工具对自己人工智能活动的可用性。

2022年，美国国家标准与技术研究院（National Institute of Standards and Technology，简称NIST）发布了一系列人工智能评估项目。这些项目涵盖生物统计、计算机视觉、鉴证、信息检索、制造业与机器人、自然语言处理、语音处理等多个类别，目前正逐步推广人工智能偏见缓解和隐私保护方面的评估项目，为人工智能产品及服务提供了有效的评估监管工具。

2022年3月，由中国电子技术标准化研究院牵头承担的“面向人工智能基础技术及应用的检验检测基础服务平台建设项目”正式上线，依托自主研制的相关标准规范，建立涵盖数据集、算法库、模型库的基础资源库和行业资源库，开展测试工具部署，搭建检验检测服务平台环境。目前，该平台正研发拓展人工智能伦理相关资源库和检验检测能力，未来将集成至“专项解决方案”模块。平台现有注册人工智能企业、科研院所300余家，致力于建成一站式人工智能检验检测公共服务平台，有效提升中国人工智能检验检测整体水平，推进人工智能技术稳步发展。

图6 人工智能检验监测公共服务平台网站

2022年5月，新加坡发布了全球首个人工智能监管测试框架和工具集（A.I. Verify）。该框架和工具集旨在融合测试和过程检查，促进企业和相关利益者之间的透明性，从而培养公众对人工智能技术的信任，同时支持人工智能技术的广泛使用。开发人员和所有者可以通过标准化测试根据原则验证人工智能系统的性能。A.I. Verify最终形成了11个关键人工智能伦理准则和5大支柱。11个伦理准则包括：透明性、可解释性、可复现性、安全性、鲁棒性、公平性、数据治理、可审计性、人类能动性与监管、包容增长、社会和环境福利。5大支柱描述了系统所有者和开发人员如何构建客户与消费者之间的信任：包括人工智能和人工智能系统使用的透明性、理解人工智能模型如何做出决策、确保人工智能系统的安全和弹性、确保公平性、确保人工智能系统的适当管理和监管。

2022年6月，西班牙政府与欧盟委员会共同发布了首个试点人工智能监管沙盒（Regulatory Sandbox on Artificial Intelligence），以推进欧盟人工智能法案（Artificial Intelligence Act）的实施。人工智能监管沙盒有效地连接了监管机构和创新者，并为双方提供可控的合作环境。监管机构和创新

者之间的这种合作促进创新人工智能系统的开发、测试和验证，并确保其符合人工智能法规的要求。监管沙盒的亮点在于有望产出易于遵循、面向未来的最佳实践指南及相关材料，以促使人工智能法规在业界（尤其是中小企业和初创企业）的有效执行。

5 人工智能伦理标准化

人工智能高质量发展离不开标准化支撑，标准也将成为人工智能伦理治理的关键要素。人工智能基础理论和技术生态发展迅速，通过研制人工智能标准，组织开展标准符合性评估评测，能够推动提升产品质量，规避潜在风险，助力人工智能产业链上下游企业之间、龙头企业与中小微企业之间协同发展，形成开放的生态体系，促进人工智能产业化规模化，引领产业健康发展，支撑社会高质量建设。

对于人工智能领域中潜在的伦理问题——从隐私和安全漏洞到歧视性结果和对人类自主性的影响，由于立法和监管往往滞后于人工智能的快速发展，亟须研制人工智能伦理相关标准，解决伦理规范和风险管理滞后于技术发展的问题，标准作为“技术法规”的效用也很好地贴合了人工智能伦理的治理需求。各方需要意识到人工智能技术的伦理和社会问题的严重性，积极参与相关标准化活动，以“优势先行、成熟先用、基础统领、应用牵引”为原则，推动人工智能伦理标准体系研制和建设，助力人工智能领域健康发展。

5.1 人工智能伦理标准化现状

从标准制定的角度看，国际上已有的人工智能标准主要是相关技术与应用的标准，而涉及人工智能伦理的相关标准仍然较少，例如伦理设计、公平性要求、风险管理等。相关指导意见多以指南和框架的形式出现，提出多种原则以规范人工智能伦理技术的使用，更加具体的标准仍处于探索和研究阶段。当前在国际标准化组织（ISO）、国际电工委员会（IEC）、国际电信联盟（ITU）等国际标准组织，以及我国国家人工智能总体组、全国信息技术标准化技术委员会-人工智能分技术委员会（TC 28/SC 42）及全国信息安全标准化技术委员会(TC 260）均已经开展了人工智能伦理相关的标准化工作（标准项目详见附表2）。

5.1.1 国际人工智能伦理标准化

ISO和IEC第一联合技术委员会（ISO/IEC JTC 1）以信息技术为核心，依托人工智能分技术委员会（SC 42）开展人工智能标准化工作。在组织架构方面，SC 42下设的可信工作组（WG 3-Trustworthiness）负责人工智能可信方面的研究和标准化工作，目前已发布3项标准，在研10项标准，是SC 42中标准项目最多的工作组；并与IT治理和IT服务管理分技术委员会（SC 40）联合成立人工智能治理工作组（JWG 1）负责人工智能治理方面工作。在国际标准项目方面，人工智能伦理写入SC 42的标准化路线图，开始从标准化预研阶段进入具体标准编制阶段，但是由于国际标准制定周期过长，目前还没有形成完善的人工智能伦理标准体系，直接相关人工智能伦理的国际标准文件还属于技术报告类型。除此之外，SC 42还针对人工智能伦理的子原则例如鲁棒性、可解释性、隐私性等开展了多项标准研制，相关国际标准项目包括但不限于：

（1）ISO/IEC TR 24368:2022《信息技术 人工智能 伦理和社会关注概述》

该标准针对人工智能系统的伦理和社会性关注内容进行梳理，从高层次概述了与人工智能系统和应用相关的道德和社会关注领域的工作架构，包括原则、流程和方法相关的信息，并在此基础上举例说明了在开发和使用人工智能过程中有关伦理和社会关注方面的实践，为企业和社会提供了伦理审查框架的指南和注意事项示例。

（2）ISO/IEC DIS 23894《信息技术 人工智能 风险管理指南》

该标准为组织在开发、生产、部署、应用人工智能技术使能的产品、系统和服务过程中提供针对人工智能的特定风险管理的指南。有关人工智能伦理的风险管理方面，该标准提及到以下几个方面---不同主体（如政府主导、监管机构、公民社会、科研机构、产业界等）应考虑针对人工智能产品设计和应用等过程中的有关伦理问题；在对于个人隐私信息的收集、

保存和使用时需要充分考虑尊重人类价值和人类尊严方面的伦理原则；具体技术细节方面也需要关注可能导致产品存在伦理问题的风险，例如在训练模型中如果使用了错误的数据或者带有偏见的数据。

（3）ISO/IEC TR 24028：2020《信息技术 人工智能 人工智能的可信赖概述》

该标准针对人工智能系统安全性、隐私性、偏差、不可预测性、实施使用阶段等9项威胁和挑战进行了研究，并提出了实现可信赖人工智能系统透明性、可解释性、可控性、无偏见、隐私保护、鲁棒性、硬件失效对策等基本原则和方法。

（4）ISO/IEC TS 6254《信息技术 人工智能 机器学习模型和人工智能系统可解释性的目标和方法》

该标准阐述了机器学习模型和人工智能系统可解释性的内涵及不同利益相关方对于可解释性的目标，列举了实现可解释的途径及人工智能系统生命周期中需要考量的有关可解释性的因素。

（5）ISO/IEC TR 24029《人工智能 神经网络鲁棒性评估》

该系列国际标准列举了目前评估神经网络鲁棒性的方法，包括统计方法、形式化方法和经验方法，以及具体实践时的使用方法。

（6）ISO/IEC TR 5469《信息技术 人工智能 功能安全与AI系统》

该标准描述了有关人工智能功能安全的特性、风险因素、控制措施及过程等，提出了实现人工智能技术应用中人身安全保护功能的方法。

国际电工委员会（IEC）自主与人工智能应用中的伦理标准评估组（IEC/ SEG 10）于2018年成立，其主要工作是识别与IEC技术活动相关的伦理问题和社会问题，并适当地向SMB（标准管理局）提出建议，为IEC委员会制定有关自主系统或人工智能应用的伦理方面的广泛适用的指导方针，确保IEC委员会之间的工作一致性，促进与ISO/IEC JTC 1/SC 42的合作等。

国际电信联盟（ITU）与40个联合国专门行政部门、瑞士政府联合建立“AI for good”（AI向善）论坛，旨在构建相互连接的平台，帮助和确定可实施的人工智能解决方案，以推进联合国可持续发展目标。该论坛的主要职能之一就是加速ITU对快速增长的重要人工智能技术进行战略研究和相关标准化工作预研。目前已经针对人工智能伦理的议题开展了多次研讨会并发布了多项研究文件，为ITU后续人工智能伦理标准的研制和发布奠定了良好的基础。

电气与电子工程师协会（IEEE）聚焦人工智能领域伦理道德标准，于2017年发布了第二版《伦理一致性设计》（Ethically Aligned Design），目的在于指导与规范人工智能在设计上合乎道德的标准，达到造福全人类的目的，避免算法偏见等潜在伦理道德风险。之后根据相关理念和准则正在研制7000系列标准（具体标准见附表2），发展人工智能应该关注设计环节，通过建立标准化的流程，组织可以在概念探索和开发的各个阶段考虑伦理价值，以获取伦理价值并确定其优先顺序，包括通过系统设计中的操作概念、价值主张、价值处置和可追溯伦理道德价值。

5.1.2 国内人工智能伦理标准化

自2018年1月18日国家人工智能标准化总体组成立之后，主要负责我国人工智能标准化统筹管理工作，最初3个专题组之一则是人工智能与社会伦理道德标准化研究组。人工智能与社会伦理道德标准化研究专题组针对人工智能标准化与伦理问题进行深入研究，于2019年4月发布了《人工智能伦理风险分析报告》，对人工智能原则进行了梳理，提出了对应的人工智能伦理评估方法。2020年，发布《人工智能伦理与社会关注国际标准研究》。工信部、国家标准委、中央网信办、发展改革委、科技部于2020年8月5日联合发布的《国家新一代人工智能标准体系建设指南》将“安全/伦理”划分为我国人工智能标准体系重要组成部分。

2020年8月，全国信标委人工智能分委会（TC 28/SC 42）下设成立可信赖研究组，重点开展人工智能系统可信赖要素的研究工作，将人工智能伦理符合性作为重要的考虑因素之一，面向人工智能系统开发全流程研究对应的评价方法和实施途径，从硬件、数据、算法等多个层面提高人工智能系统的可信赖能力。并在2022年发布了《人工智能可信赖标准化白皮书》，将可信赖人工智能分为功能安全性和伦理符合性两大基本原则并衍生出15条子指标，例如可靠性、可问责性、可解释性等，系统描述出可信赖人工智能需解决的伦理问题，以及相应的技术框架和流程框架，为后续相关标准研制提供了坚实的理论基础。另外，SC 42发布T/CESA 1193—2022《信息技术 人工智能 风险管理能力评估》，其中规定了人工智能产品的风险管理能力评估体系及评估流程，提出了多项关于人工智能伦理的规范，并要求组织成立伦理委员会对人工智能企业的伦理风险进行专门管理。目前，SC 42正在开展风险评估、隐私保护、可信赖规范等国家标准研制，积极推动人工智能伦理相关标准立项。

全国信息安全标准化委员会（TC260）负责组织开展国内信息安全有关的标准化技术工作，2021年1月发布了《网络安全标准实践指南—人工智能伦理安全风险防范指引》，该指引依据法律法规要求及社会价值观，针对人工智能伦理安全风险，给出了安全风险防范措施，为相关组织或个人在各领域开展人工智能研究开发、设计制造、部署应用等活动时提供指引。

5.2 人工智能伦理标准体系

为了贯彻落实相关国家政策文件，结合人工智能伦理技术及标准化研究现状，形成人工智能伦理标准体系结构，围绕人工智能伦理技术研究及应用，规范人工智能服务冲击传统道德伦理和法律秩序而产生的要求，充分发挥标准的引领作用，面向市场和技术发展需求逐步开展相关标准化工作，为人工智能普及应用和创新发展做好保障，努力实现人工智能高质量

发展与高水平治理的平衡。同时，做好总体设计和布局，加强关键技术域标准研制，形成系列协调配套的关键标准，提升我国人工智能伦理标准的先进性和国际影响力，形成标准引领人工智能产业健康发展的新格局。

人工智能伦理标准体系结构包括“基础共性”“治理技术”“管理”“行业应用”4个部分，如下图所示：

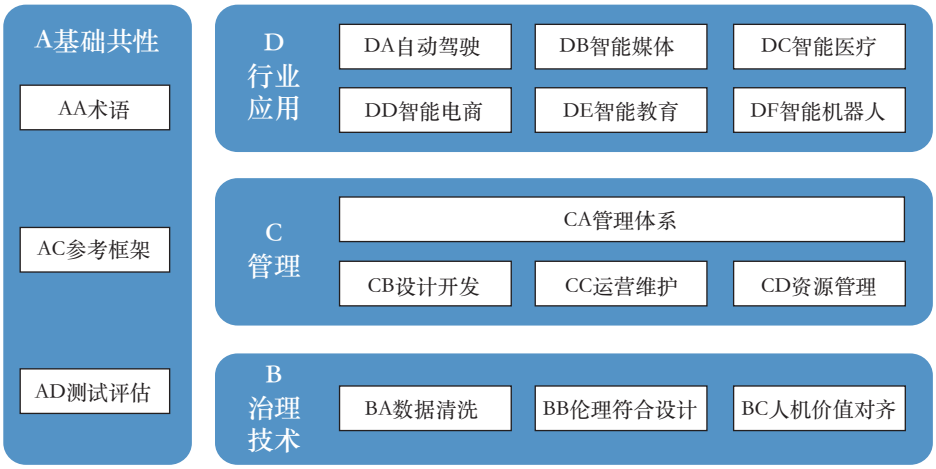


图7 人工智能伦理治理标准体系结构图

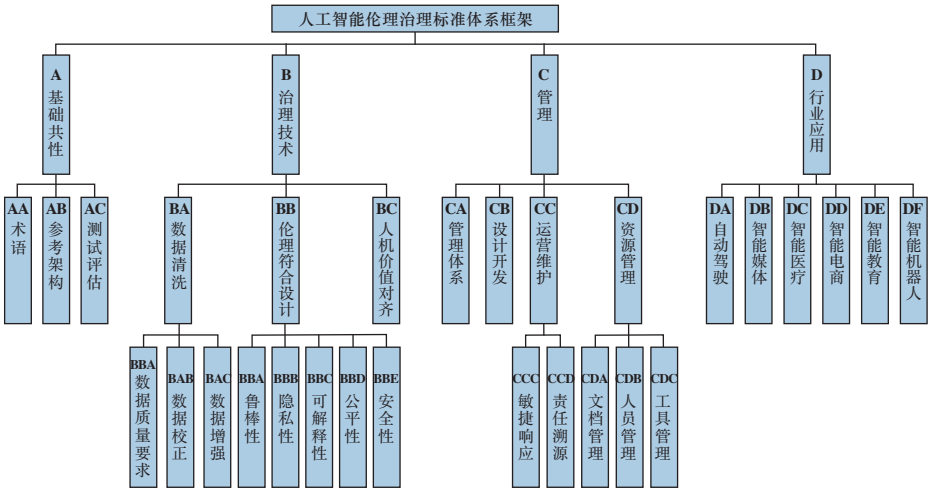


图8 人工智能伦理治理标准体系框架图

5.2.1 A基础共性标准

该部分为体系结构中的其他部分提供支撑，有助于为各界建立统一的人工智能伦理基础，其中：

“AA术语”类标准确立整体术语，定义和描述其概念的内涵和外延，帮助各利益相关方更好地理解人工智能伦理治理的基本情况和涉及范围；

“AB参考架构”类标准提供人工智能伦理技术框架、管理办法等方面的逻辑关系和相互作用，促进产业形成全面的人工智能伦理治理体系；

“AC测试评估”类标准提供面向人工智能伦理要求的测试方法理论和评估指标，制定符合具体实践要求的测试评价标准，支撑建立人工智能伦理合规体系，保障人工智能产业健康发展。

5.2.2 B治理技术标准

该部分旨在规范人工智能伦理治理的支撑技术，加强对伦理问题事前控制的可操作性和规范性，辅助人工智能伦理治理落地。其中：

“BA数据清洗”类标准针对训练数据集、测试数据集等关键数据环节，包括针对人工智能伦理方向的数据集质量要求，以及缓解人工智能伦理问题的数据处理方法（根据实现方法不同，相关技术可以分成“数据校正”、“数据增强”等）；

“BB伦理符合设计”类标准围绕隐私性、安全性、鲁棒性、公平性、可解释性等重点伦理准则，分别提出标准化的伦理符合设计技术或衍生工具，将抽象的伦理准则映射到实际的方法论，提高人工智能伦理治理的预前能力，帮助人工智能产业在应用产品中落实关键人工智能伦理准则；

“BC人机价值对齐”类标准旨在将人类价值观通过编码、交互、预学习等方式嵌入人工智能应用，该类技术标准有助于人工智能应用更好地满足用户需求，也有助于实现强人工智能或通用人工智能，特别对于数字人、智能客服、类人机器人等仿人智能产品。

5.2.3 C管理标准

该部分包含组织中为保障人工智能伦理要求，所需要协调统一的管理事项的标准，提升人工智能解决方案全流程的生产力。其中：

“CA管理体系”类标准，考虑组织管理涉及的关键过程，为组织规范使用和管理人工智能提供框架、行动指南和风险管理要求，指导人工智能系统提供方、使用方符合正确伦理观地进行开发或使用人工智能系统。

“CB设计开发”类标准面向人工智能产品应用生命周期的设计开发阶段，通过管理手段保障内部流程有充足能力，将合规的技术和相应的工具使用到产品设计开发过程中，以确保合理的伦理准则要求得到满足，将治理点提前到产品的研发阶段，确保产品在应用和市场化之间规避伦理风险。

“CC运营维护”类标准面向人工智能产品应用生命周期的部署服务阶段，主要包括：针对不可预知的人工智能伦理问题，确保合理的敏捷响应机制能够迅速明确问题，并具有足够的技术储备快速解决问题；根据人工智能伦理相关问题的责任的特点，制定明确的责任溯源机制和方法。

“CD资源管理”类标准，根据资源种类不同，具体分为“文档管理”（文档的编写、日志的记录和存储维护规范）“人员管理”（面向人工智能管理者、开发者、使用者等不同群体，提出对应的从业人员要求和人力资源管理要求）“工具管理”（为保障人工智能系统的伦理符合性，对相关开发、测试等工具进行合规管理，确保在各时间点具备合适的工具支持）。

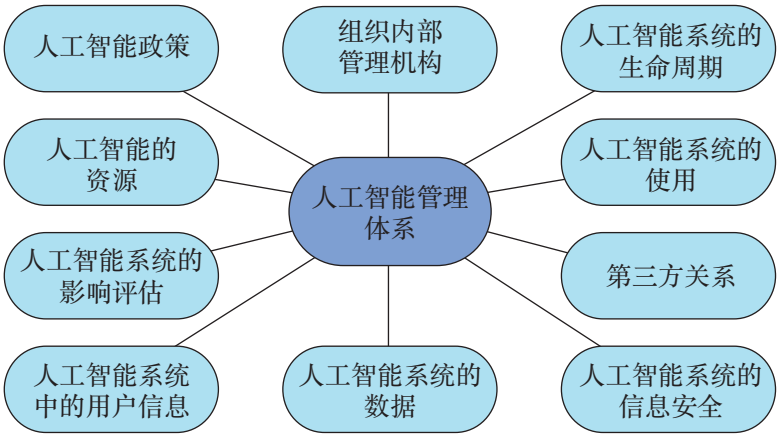
5.2.4 D行业应用标准

该部分重点围绕DA自动驾驶、DB智能媒体、DC智能医疗、DD智能电商、DE智能教育、DF智能机器人等人工智能伦理高风险高敏感领域，开展“急用先行”的标准研究，深度挖掘各智能领域在人工智能伦理治理方面的标准化需求和标准化方法，引领相关产业发展。由于人工智能伦理

准则的内涵和具体要求会随着应用场景而改变，因此需要结合各行业应用现状，提出符合特定应用场景的要求，保障人工智能伦理治理前瞻性、适用性、及时性，有效提升人工智能广泛赋能传统行业下的用户体验，提高智能化应用在人工智能伦理方面的规范化水平，为有关政府部门提供监管抓手。

5.3 重点标准研制

5.3.1 人工智能 管理体系



《人工智能 管理体系》（国家标准计划号：20221791-T-469）规定了人工智能应用的组织政策，组织部门架构及其相关职责，负责任使用人工智能的过程及目标，组织与第三方之间的职责与关系；

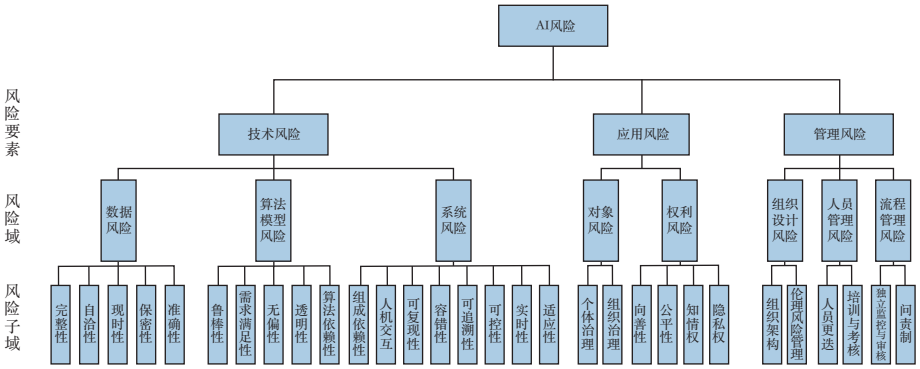
提供针对人工智能应用涉及的资源安排与管理办法、组织使用人工智能的评估过程及对组织内外部个人、群体及社会影响；

提出组织在基于人工智能生命周期的调研与分析、设计与开发、验证与确认、部署、运行与监测、重新评估及退出阶段中可实现负责任人工智能应用的流程及目标；

确定人工智能应用中的数据质量管理规范和保证信息安全的策略与措施。

5.3.2 人工智能 风险评估模型

《人工智能 风险评估模型》规定了人工智能领域产品的风险评估模型，包括风险能力等级、风险要素、风险能力要求，给出了判定人工智能产品的风险评估等级的方法。人工智能风险评估模型整体框架如图9所示。该标准可以指导人工智能产品开发方、用户方以及第三方等相关组织对人工智能产品风险开展评估工作。



5.3.3 人工智能 隐私保护机器学习技术要求

《人工智能 隐私保护机器学习技术要求》规定了隐私保护机器学习系统的技术要求，来规范化隐私保护机器学习系统的技术框架及流程、功能要求、非功能性要求和安全要求。该标准用于指导科技企业、用户机构、第三方机构等，对隐私保护机器学习系统的设计、开发、测试、使用、运维管理等。

6 人工智能伦理治理的展望与建议

人工智能作为新一轮科技革命和产业变革的重要驱动性技术，推动经济社会各领域从数字化、网络化向智能化加速跃升。但是，随着人工智能赋能的日益广泛，泄露个人隐私、冲击就业格局、危害公共安全等风险问题也在不断扩大，给社会治理带来全新挑战。在今日的智能化社会，人工智能无处不在，代替人做出各种各样的决策，潜在的伦理问题也逐渐显现。为了应对人工智能应用对传统科技伦理观的强烈冲击以及人工智能技术对人文社会带来的巨大转变，人工智能各利益相关方需要秉持科技造福人类、安全与发展兼顾的治理目标，平衡好创新发展与有效治理的关系，提出切实有效的治理举措，根据实际需求及反馈动态调整、持续优化治理路径及策略，携手共建全球人工智能伦理治理体系。

6.1 动态更新人工智能伦理准则，推动凝聚国际共识

据统计，目前全球已经有160多个国家出台了人工智能的伦理原则或指南，相关人工智能伦理准则已经超过40项。从发布内容上看，除文化、地区、经济等因素引起的差异外，目前的人工智能伦理准则已形成了一定的社会共识，尤其是在以人为本、促进创新、保障安全、保护隐私、明晰责任等伦理观上取得了高度共识，但仍有待继续加深理论研究和论证，努力尝试在世界范围内建立新的共识。为使伦理准则能够在具体的应用场景中落实，需要进一步制定有针对性、时效性、可操作性的实施细则。

因此，未来对于人工智能伦理准则还需要朝着以下两点目标发展。一是为了改善目前人工智能伦理准则过于宏观、不易实操落地问题，需要统一“政产学研用”各界及不同国家地区、不同人群等对于人工智能伦理的看法和要求，进一步明确人工智能伦理观的范围，丰富其内涵，拓展其外延，对现有准则进行完善和细化，发展更“接地气”的人工智能伦理准则。二是秉着“求同存异”的理念，在国际共识的基础上，融入我国传统

文化价值观和人工智能产业发展现状，发展有中国特色的人工智能伦理准则，适应我国人工智能发展的特殊需求，更好地帮助我国各界规范人工智能伦理问题。

（1）加强国际交流合作，积极达成全球共识。在发扬中国特色的同时兼具国际视野，考虑全球宗教、传统文化、价值观等因素，以及各学派在传统科学、社会科学、哲学等观点的差异性，通过学术会议、国际标准研讨会议等国际会议平台寻求共识，积极争取在国家或区域之间构建人工智能伦理治理准则合作备忘录，求同存异，令准则充分满足不同人群的需求，以人类命运共同体的观念治理人工智能技术带来的伦理风险。

（2）明确准则内涵外延，形成文理跨学科合力。不断细化伦理准则的内涵外延，广泛邀请多学科、跨学科专家参与讨论，建立领域知识互认、统一术语概念等合作机制，使治理工作在跨领域协作时能统一范畴，形成合力。一方面，让哲学、社会学、法学等学者充分了解目前人工智能技术的发展现状和局限性，避免“空中楼阁”问题，即过分探讨太过遥远的问题，或者担忧不会发生的问题，防止一定程度上阻碍技术发展。另一方面，让工程学、数学、机器人学等学者充分了解社会对人工智能技术的担忧，以及对人类社会的潜在影响，从而将社会性问题的考虑充分融入技术的设计研发过程，而不是单一追求功能或性能的突破。

（3）研究准则动态更新方式方法，适应技术快速演变。不同于成熟的医学伦理学，从只精通下围棋的任务型AI-AlphaGo到能够在多专业领域持续输出知识甚至独立观点的ChatGPT，人工智能技术加速演进，能够提供的功能也不断进化，相匹配的准则要求也不断变化，需要动态且敏捷的更新。根据具体功能优化对技术的限制要求，才能适应人工智能技术的频繁换代。同时，强化对人工智能立法的支撑作用，与人工智能立法协作配套，为立法提供理论基础，并在立法要求上持续完善补充，提供法治到公众之间的桥梁。

6.2 加速打造多方协同的治理模式，促进政产学研用治理深度融合

人工智能治理的重要特征之一是治理主体的多元化，因此构建持续发展的人工智能治理体系需要包括政府、企业、大学和科研机构、标准化组织、行业组织、政府机构到法律部门等在内的多方合作，各司其职、各尽其能、相互协调和协同共治，以人工智能伦理准则为指导开发可靠、稳健、可信、安全和可协作的人工智能系统。

为了应对人工智能技术的特殊性，“技术+制度”的治理体系要求人工智能治理在充分发挥技术手段的基础上，注重伦理与制度的结合，逐步形成“标准引领+技术防范+法律规制”的综合规制体系。从近期看，着重加强伦理驱动的技术研究和标准规范。遵循市场规律，坚持应用导向，以可实践性作为主要目标，完善人工智能标准体系，将隐私保护、风险管理和可信赖规范等作为我国新一代人工智能标准体系中“伦理/安全”部分的重要发力方向；从中长期考虑，全面研究和论证人工智能法律规制体系，制定立法策略。围绕国家和社会层面对人工智能领域立法的需求，在人工智能发展相对成熟的领域，适时开展相关的规范立法形成有效的归因机制，解决人工智能新场景下的责任划分问题；同时需要推动建立细分领域的法律法规体系，形成行之有效的监管和治理机制，用法律手段守护人工智能应用的“底线”。

（1）构建“政产学研用”连通的治理机制。构建贯通多维度的合作机制和数据流通机制，保证信息数据通畅，解决伦理治理相关知识和问题（“数据孤岛”“数据烟囱”等问题）互通不及时造成的治理滞后性。其中包括：协调机制（帮助不同社会角色的参与）、决策咨询机制（为重大决策提供支撑证据）、参与和对话机制（政府和专家之外的利益相关者和公众参与决策与管理的机制）、监测评估与动态调整机制（把握政策实施的方向、进展和问题，提出改进和调整的意见）、审查与监管机制（通过可实施的审查规范落实相关政策法规并持续对市场监管）。

（2）建设人工智能伦理治理创新示范区。在人工智能科技公司高密度地区建设伦理治理的创新示范区，呼吁有能力的企业、科研院所投入示范区的建设。在一定范围内验证伦理治理的标准规范、治理工具、协同工作和数据流通共享机制等治理手段的科学性，为大范围普及应用或市场准入门槛的设定提供依据，促进标准的落地实施。示范区的经验和案例，可以进一步完善和修订相关标准，促进伦理治理标准的迭代升级。在科技创新方面，围绕可持续发展（如绿色低碳、可循环利用等）和科技向善（如信息无障碍、残障人士帮助、适老化等方面）等非营利性方向加大研发力度，利用资金、政策支持等方式促进企业加大相关投入力度。

（3）建立公共技术服务或检验检测平台。通过高质量公共平台有效帮助中小企业合规发展，整体提高产业水准，拉齐人工智能行业的伦理“基准线”。一是提供基础技术服务。提供训练数据集、基准数据集、偏见缓解、隐私保护、鲁棒性等关键治理技术元素，帮助企业快速提升自身治理水平，并体现在产品和服务中。二是提供评测认证服务。提供方便快捷的检验检测服务和一体化认证流程，整体拉高市场准入水平，将伦理风险扼杀在进入市场前；同时提高效率、降低成本，减轻对企业发展造成的限制作用。三是持续监测。完善公共平台后台数据管理，一方面是人工智能企业、技术、产品等方面的动态监测机制；另一方面是用户、消费者、受技术影响者等社会相关方的举报申诉通道，及时发现市场中治理手段实施中的问题，促进标准等治理抓手的持续改进和优化。

6.3 强化支撑技术的实践水平，跨越准则到治理技术鸿沟

提升人工智能伦理治理水平，治理技术的创新研究必不可少。特别在基础理论研究、隐私计算、价值观嵌入等重点方面，加强技术研究力度，让伦理保障技术与人工智能应用技术发展速度相匹配，例如：人工智能工程设计哲学与设计伦理将技术创新方法“发明问题解决理论”(TRIZ)引入

现代技术哲学研究，以现代技术的“工程设计”及其结果“人工智能”为研究对象进行哲学和伦理反思；隐私设计和隐私计算技术在系统设计的最初阶段将个人信息保护的需求嵌入其中，成为系统运行的默认规则，配合加密机制在保证原始数据安全和隐私性的同时，完成对数据的计算和分析任务等。

针对人工智能伦理实践，要同步通过改进技术或开发技术工具来将创新成果转化为技术解决方案，例如开发贯穿整个人工智能生命周期的技术工具、管理工具，以及按照应用场景将需求特性集成到人工智能系统中的工具包（Toolkit）。同时，由于人工智能技术更迭快、创新发展迅速，为实现更为敏捷的治理，可适当引入治理效果评价及反馈等机制，构建覆盖全面、反馈及时的治理效果评价机制，以准确掌握最新治理情况、及时发现问题，以制定更有效的应对策略，例如伦理准则的可操作性评估、技术工具的安全性评价等。

（1）深耕治理底层理论和保障性技术研究。突破符号学习、知识计算等具有可解释性的底层机器学习范式，促进人工智能技术“去黑盒化”。针对人工智能伦理准则部署，攻关隐私计算、联邦学习等底层隐私保护技术；对抗学习、多任务学习等鲁棒性增强技术；公平性预处理、公平性判别器等公平性增强技术等等一系列关键准则对应的治理技术，跨越伦理准则到治理技术之间的鸿沟。同时，积极将数据处理技术、伦理符合性设计、人机价值对齐等先进的技术转化成技术标准或工具，更好地普及应用。

（2）面向重点敏感领域攻关专项治理技术。针对自动驾驶、智能医疗、脑机接口、智能教育等高风险行业领域，开展专项课题研究，建设治理专用数据集，提升高风险领域的治理技术水平。围绕领域技术和应用场景特点，一是探索治理技术的“捷径”，将伦理风险限制在技术投产之前，即通过设计方法或嵌入技术将伦理要求植入人工智能产品的技术路

线；二是研究合规性测试评估方法，高效测试人工智能产品的各项能力，对市场准入门槛进行合理规范。

（3）采用先进技术孵化配套开源实施工具。通过研发工具将先进技术投入实际应用，通过开源的方式共建生态，整体提升人工智能产业的伦理合规性。开源通过公开透明的方式降低边际成本，使高复杂度的软件工具人人可贡献、人人可用、人人可维护，帮助开发公司降低成本，帮助中小企业“搭便车”。基于开源项目的代码公开、规则公开、过程公开的开发环境，组建技术委员会，促进行业内交流和共建，推动以协作的方式解决行业共性问题。开源通过社区协作机制进行激发技术创新，降本增效的同时加速技术发展和扩大影响范围。

6.4 强化标准布局，引领产业高质量发展

全面建成人工智能伦理治理体系，标准化是不可或缺的关键环节。将标准作为准则和实践之间的桥梁，针对不同应用场景做出对应的规定，深化标准的对技术的规范作用。建设人工智能伦理标准体系，首先要重视重点领域的推荐性标准制定工作，在智能医疗、自动驾驶、智能教育等伦理高风险领域，制定一系列包括产品标准、技术标准、管理标准等在内的具体要求和规范，发挥急用先行标准的引领作用。我国在人工智能领域具备一定的优势，需要更加积极主动地应对人工智能伦理问题带来的挑战，在人工智能标准化中承担相应的前瞻研究，例如人工智能应用的伦理风险分类分级标准等。并推广基于标准的认证认可，围绕人工智能伦理的特殊要求，对人工智能企业的管理能力进行审计，加强人工智能产品与服务投入使用前的评估，充分保障各利益相关方的合法权益。

（1）在相关技术标准方面。规范人工智能伦理治理的支撑技术，从数据处理技术、伦理符合性设计、人机价值对齐等方向加强事前控制和风险预防的可操作性和规范性，辅助人工智能伦理治理落地。用标准固化产

业内成熟的治理技术，通过标准符合性测试、培训等方式宣贯实施，促进更多的利益相关方熟悉如何用标准化的技术手段提高自身治理水平，提供产业内人工智能伦理治理的技术性抓手，从开发的源头上提高人工智能产品的伦理合规性。

（2）在相关认证标准方面。发布一批人工智能伦理符合性认证标准，一是产品认证，通过测试、评估、审计等方式对人工智能技术应用及产品进行分类分级；二是组织认证，规范组织中为保障人工智能伦理要求，所需要协调统一的资源管理、内部流程、组织架构等。通过标准引导产业内企业在技术研发和产品设计中落实人工智能伦理治理要求，建立起人工智能伦理治理至人工智能技术之间的桥梁。通过认证的形式引导产业内企业将人工智能伦理治理融入日常研发生产。

（3）在高风险领域标准方面。面向自动驾驶、教育、医疗等关键技术应用领域，针对不同行业特点和风险类型，制定行业伦理规范、行为准则、针对性治理技术和从业人员要求等标准，通过行业组织、教育培训、人才认定等方式，加大伦理规范的执行力度，促进伦理要求落地。并从技术层面规范关键技术应用的伦理风险要点，提高关键技术应用的伦理治理水平，防范伦理风险，进一步提高人工智能伦理治理的针对性和实效性。

在国际标准化方面，积极开展国际对话与合作，在充分尊重各国人工智能治理原则和实践的前提下，主导相关国际标准的制定，把握科技发展话语权，引领国际人工智能伦理治理方向。当前，人类社会正步入智能时代，世界范围内人工智能领域的规则秩序处于形成期，需要在最具代表性和突破性的方面中占据标准化制高点，为实现人工智能的全球治理作出积极贡献，推动形成具有广泛共识的国际人工智能治理框架和标准规范。

附件 1

标准体系明细表

一级	二级	标准号/计划号	标准名称	状态
A基础共性	AA 术语	GB/T 41867-2022	信息技术 人工智能 术语	发布
	AB 参考架构	20203869-T-469	人工智能 面向机器学习的系统规范	国标在研
	AC 测试评估	20221450-T-469	人工智能 深度学习算法评估	国标在研
		20221348-T-469	人工智能 服务能力成熟度评估	国标在研
B治理技术	BA 数据清洗	20201611-T-469	人工智能 面向机器学习的数据标注规程	国标在研
		20220787-T-469	信息安全技术 网络数据分类分级要求	国标在研
		GB/T 41574-2022	信息技术 安全技术 公有云中个人信息保护实践指南	发布
		GB/T 41871-2022	信息安全技术 汽车数据处理安全要求	发布
	BB 伦理符合设计	——	人工智能 联邦学习技术规范	国标拟立项
		——	人工智能 隐私保护机器学习系统技术要求	拟研制
C 管理	CA 管理体系	20221791-T-469	人工智能 管理体系	国标在研
	CB 设计开发	——	人工智能 可信赖规范 第1部分：通则	团标在研
		——	人工智能 可信赖规范 第2部分：计算设备	团标在研
		——	人工智能 可信赖规范 第3部分：机器学习框架	团标在研
		——	人工智能 可信赖规范 第4部分：机器学习模型	团标在研
	CD 资源管理	——	人工智能 风险管理能力评估	国标拟立项
D 行业应用	DA 自动驾驶	——	人工智能 自动驾驶系统仿真测试平台技术要求	国标拟立项
		——	人工智能 可信赖规范 第6部分：自动驾驶	拟研制
	DB 智能媒体	2022-1322T-SJ	人工智能 深度合成图像系统技术规范	行标在研
		2022-1323T-SJ	人工智能 视频图像内容审核系统技术规范	行标在研
	DD 智能电商	GB/T 40094.4-2021	电子商务数据交易 第4部分：隐私保护规范	发布
	DF 智能机器人	2022-1321T-SJ	人工智能 计算机视觉系统可信赖技术规范	行标在研
		——	人工智能 家庭智能中枢系统可信赖技术规范	团标在研

附件 2

人工智能伦理相关国际标准清单(截止2022年底)

序号	标准归口单位	标准编号	标准化文件名称	状态
1	ISO/IEC JTC1/SC42	ISO/IEC AWI TS 6254	《信息技术 人工智能 ML模型和AI系统可解释性的目标和方法》 Information technology — Artificial intelligence — Objectives and approaches for explainability of ML models and AI systems	在研(AWI)
2	ISO/IEC JTC1/SC42	ISO/IEC AWI TS 8200	《信息技术 人工智能 自动化人工智能系统可控性》 Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems	在研(AWI)
3	ISO/IEC JTC1/SC42	ISO / IEC DIS 23894	《信息技术 人工智能 风险管理指南》 Information Technology — Artificial Intelligence — Guidance on risk management	在研(DIS)
4	ISO/IEC JTC1/SC42	ISO / IEC TR 24027：2021	《信息技术 人工智能 AI系统和AI辅助决策的偏见》 Information technology — Artificial Intelligence — Bias in AI systems and AI aided decision making	发布
5	ISO/IEC JTC1/SC42	ISO / IEC TR 24028：2020	《信息技术 人工智能 可信赖人工智能概述》 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence	发布
6	ISO/IEC JTC1/SC42	ISO / IEC TR 24029-1:2021	《人工智能 神经网络鲁棒性评估 第1部分：概述》 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview	发布
7	ISO/IEC JTC1/SC42	ISO / IEC CD 24029-2	《人工智能 神经网络鲁棒性评估 第2部分：使用形式化方法的原则》 Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods	在研(CD)
8	ISO/IEC JTC1/SC42	ISO / IEC DTR 24368	《信息技术 人工智能 伦理和社会问题概述》 Information technology — Artificial intelligence — Overview of ethical and societal concerns	发布
9	ISO/IEC JTC1/SC42	ISO/IEC PWI TS 12791	《信息技术 人工智能 分类中非预期偏见的处理方法》 Information technology – Artificial intelligence - Treatment of unwanted bias in classification	在研(PWI)
10	ISO/IEC JTC1/SC42	ISO/IEC 38507:2022	《信息技术 IT治理 组织使用人工智能的治理影响》 Information technology - Governance of IT - Governance implications of the use of artificial intelligence by organizations	发布
11	ISO/IEC JTC1/SC42	ISO/IEC 42001	《信息技术 人工智能 管理体系》 Information technology - Artificial intelligence - Management System	在研(CD)
12	IEEE	-	《人工智能设计的伦理准则（第2版）》 Ethically Aligned Design (Version 2)	发布

序号	标准归口单位	标准编号	标准化文件名称	状态
13	IEEE	IEEE P2894	《可解释人工智能的结构框架指南》 Guide for an Architectural Framework for Explainable Artificial Intelligence	在研
14	IEEE	IEEE P2976	《XAI标准 可解释人工智能 人工智能系统设计的透明度和互操作性》 Standard for XAI – eXplainable Artificial Intelligence - for Achieving and Interoperability of AI Systems Design	在研
15	IEEE	IEEE 7000-2021	《系统设计期间解决伦理问题的建模过程》 Model Process for Addressing Ethical Concerns During System Design	发布
16	IEEE	IEEE 7001-2021	《自动化系统透明度标准》 Standard for Transparency of Autonomous Systems	发布
17	IEEE	IEEE 7002-2022	《数据隐私处理过程标准》 Standard for Data Privacy Process	发布
18	IEEE	IEEE P7003	《算法性偏见考量》 Algorithmic Bias Considerations	在研
19	IEEE	IEEE P7004	《儿童和学生数据治理标准》 Standard for Child and Student Data Governance	在研
20	IEEE	IEEE P7005	《透明雇员数据治理标准》 Standard for Transparent Employer Data Governance	发布
21	IEEE	IEEE 7007-2021	《伦理驱动的机器人和自动化系统的存在论标准》 Ontological Standard for Ethically Driven Robotics and Automation Systems	发布
22	IEEE	IEEE P7008	《推动伦理驱动的机器人、智能和自动化系统的标准》 Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems	在研
23	IEEE	IEEE P7009	《自动化和半自动化系统的失效安全标准》 Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems	在研
24	IEEE	IEEE 7010-2020	《评估自动和智能系统对人类福祉影响的推荐性实践》 Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-being	发布
25	IEEE	IEEE P7011	《新闻来源可信度的鉴定和评级过程的标准》 Standard for the Process of Identifying and Rating the Trustworthiness of News Sources	在研
26	IEEE	IEEE P7012	《机器可读的个人隐私条款的标准》 Standard for Machine Readable Personal Privacy Terms	在研
27	IEEE	IEEE P7014	《自动和智能系统中拟真同理心的伦理因素的标准》 Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems	在研
28	IEEE	IEEE P7015	《数据和人工智能读写能力、技能和成熟度的标准》 Standard for Data and Artificial Intelligence (AI) Literacy, Skills, and Readiness	在研

附件 3

人工智能评估评测工具清单

治理用途	工具名称	所属公司	简要介绍
偏见检测	Audit-AI	Pymetrics	测量和减轻训练数据中歧视性模式，以及为社会敏感决策过程而训练的机器学习算法所做的预测的影响
	Watson OpenScale	IBM	检测和降低AI决策的实时偏见
	AI fairness 360	IBM	检测和降低非监督模型的偏见问题
公平性	PAIR	谷歌	用户可通过What-If可视化工具体验评估五种不同的公平模式，以回答人工智能系统中提出的最困难，最复杂，最完全人性化的问题之一：用户希望将什么视为公平?
	Fairness tool	埃森哲	用于减少AI程序的种族与性别歧视
	Fairlearn	微软	一个Python工具包，使人工智能系统的开发人员能够评估其系统的公平性并减轻任何观察到的不公平问题
隐私保护	TensorFlow Privacy	谷歌	评估机器学习模型的隐私特性
	MindArmour	华为	评估人工智能模型的隐私性能
可解释性	AI Explainability 360	IBM	以可扩展的开源工具包形式帮助用户了解机器学习模型如何在整个 AI 应用程序生命周期中通过各种方式预测标签
	InterpretML	微软	训练玻璃箱模型的可解释性及解释黑箱系统，帮助用户了解模型的全局行为，或了解单个预测背后的原因
鲁棒性	RobustART	京东等	人工智能模型鲁棒性评测基准
	Cleverhans	谷歌	对抗鲁棒性机器学习库
安全性	PaddleSleeve	百度	评估人工智能模型的安全与隐私性能，提供模型加固和隐私增强手段
	MindArmour	华为	评估人工智能模型的安全性能

中国电子技术标准化研究院

电话：010-64102854
邮箱：sc42ai@cesi.cn
地址：北京市东城区安定门东大街 1 号

全国信标委
人工智能分委会

国家人工智能
标准化总体组