《人工智能伦理风险的治理要求》地方标准编制说明

一、工作简况

（一）任务来源

2020年12月21日，山东省市场监督管理局印发《2020年第二批地方标准制修订计划项目》的通知（鲁市监标字〔2020〕329号），本标准列为推荐性地方标准，计划编号为“2020年第二批地方标准制修订计划项目—12”，立项标准名称为《人工智能伦理风险的治理要求》。本文件由山东省工业和信息化厅提出，由山东省人工智能标准化技术委员会归口。

（二）起草单位、起草人及任务分工

本标准起草单位包括山东省计算中心（国家超级计算济南中心）、齐鲁工业大学、山东新一代标准化研究院有限公司、聊城大学、山东科技大学、山东省人工智能研究院、神思电子技术股份有限公司、山东博远视讯信息技术有限公司、山东省人工智能产业联盟、山东博物馆。

本标准的主要起草人为高永超、钱恒、吴林、卢晓建、周茹、井焜、曾庆田、贾仰理、单珂、翟虎、李超、苏冠群、蒋金广。所做工作如下：

高永超、钱恒负责总体设计和技术把关，吴林、卢晓建、周茹负责标准调研、文本起草，周茹负责技术资料收集分析，井焜、

曾庆田、贾仰理、单珂、翟虎、李超、苏冠群、蒋金广等参与标准起草。

（三）起草过程

1.成立起草组

山东省市场监督管理局标准制修订计划下达后，山东省计算中心（国家超级计算济南中心）高度重视本标准的制定工作，为确保标准制定的科学性、普适性和严谨性，于2021年1月成立标准起草工作组，负责本标准的制定工作。

2.形成标准工作组讨论稿

标准起草工作组充分收集研究国际国内有关人工智能伦理风险的文献资料，重点了解人工智能伦理风险的类别以及造成人工智能伦理风险的主要原因。在此基础上，标准起草工作组认真研究了国家及山东省关于人工智能伦理的政策文件，经过多次讨论和修改形成标准组讨论稿。

3.形成标准征求意见稿

2022年9月，在山东省人工智能标委会内部征求本标准草案的修改意见，标准起草工作组从内容科学性、表述规范性等方面，对标准文本进行了相应的修改完善，形成本标准的征求意见稿。

二、地方标准制定目的和意义

以人工智能技术的研发和应用为代表的新一轮科技革命，在

方便人们生活的同时也带来了复杂的伦理问题和治理挑战，人工智能伦理已经成为各国各地区持续关注的重点。2020 年国家标准化管理委员会、中央网信办、国家发展改革委、科技部及工业和信息化部联合印发《国家新一代人工智能标准体系建设指南》，将人工智能伦理列入国家人工智能标准体系，提出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。2022 年中共中央办公厅、国务院办公厅印发《关于加强科技伦理治理的意见》，首次对我国科技伦理治理工作作出系统部署，填补了我国科技伦理治理的制度空白。

为解决人工智能技术给个人或社会带来的伦理风险，制定本标准。本标准的制定填补了国内有关人工智能伦理治理的标准空白，可减少人工智能技术带来的负面影响，有助于推动人工智能技术在各领域的广泛应用，也有利于人工智能安全、可靠、可控发展。

三、地方标准编制原则、主要技术内容和确定依据

（一）标准编制原则

1.科学性原则

本标准在遵循国家关于人工智能相关政策技术文件的基础上开展，在制定过程中参考了国内外有关人工智能伦理风险类别及成因等技术文献，并及时根据相关专家反馈的问题进行调整完善，合理确定了人工智能伦理风险各方面的治理要求，具有较高

的科学性。

2. 可操作性原则

本标准制定过程中，起草工作组对人工智能各利益相关者进行了深入的研究，力求标准内容全面具体，给出了对数据、研发者、用户等的具体要求，便于人工智能利益相关者在实践中开展人工智能伦理风险的治理工作，具有较高的可操作性。

3. 规范性原则

本标准依据 GB/T 1.1—2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》的规定起草，符合标准编写要求。

（二）标准主要内容

（1）范围

本章给出了本标准的章节内容及适用范围。本标准针对人工智能伦理风险的不同来源，提出了对数据、研发者、生产商、用户的治理要求，以及对人工智能利益相关组织的基本治理要求；适用于研发和使用人工智能产品和服务的组织针对人工智能伦理风险的治理。

（2）规范性引用文件

本章给出了本标准规范性引用的文件。

（3）术语和定义

本章给出了适用于本标准的术语和定义。

（4）人工智能伦理风险类别

本章给出了人工智能存在的伦理风险类别，包括算法歧视、算法不可解释、算法决策困境、算法安全风险、算法滥用、隐私泄露、算法责任。

（5）治理要求

本章给出了对数据、研发者、人工智能产品生产商及用户的人工智能伦理风险治理要求，并提出了对人工智能利益相关组织的基本治理要求。

（三）确定依据

1.术语和定义的确定

现行国家标准 GB/T 5271.28-2001《信息技术 词汇 第 28 部分:人工智能 基本概念与专家系统》中有对“人工智能”的定义（GB/T 5271.28-2001，28.01.02 人工智能），该定义易于理解，同时基于优先采用现行国家标准的原则，本标准中使用的“人工智能”的定义，直接引用了 GB/T 5271.28-2001 的 28.01.02。

现行国家标准 GB/T 38736-2020《人类生物样本保藏伦理要求》中有对“伦理”的定义（GB/T 38736-2020，3.11 伦理），基于优先采用现行国家标准的原则，本标准中使用的“伦理”的定义，直接引用了 GB/T 38736-2020 的 3.11。

ISO 31000:2018《风险管理指南》中有对“风险”的定义（ISO 31000:2018,3.1），本标准中使用的“风险”的定义，直接引用了 ISO 31000:2018 的 3.1。

2.人工智能伦理风险类别的确定

国际标准 ISO/IEC TR 24368:2022《信息技术 人工智能 伦理和社会问题概述》阐述了人工智能技术面临的伦理和社会问题以及人工智能应遵循的关键原则。国家人工智能标准化总体组 2019 年发布的《人工智能伦理风险分析报告》将人工智能技术的伦理风险划分为算法相关的伦理风险、数据相关的伦理风险、应用相关的伦理风险、长期和间接的伦理风险四大类。本标准在参考国内外相关标准及文献的基础上，结合我国人工智能发展以及道德伦理的国情，给出了人工智能伦理风险的不同类别。

3.治理要求的确定

通过查阅国内外标准及各类文献发现，人工智能的利益相关者可以分为四类，包括提供用于训练人工智能的数据的组织、人工智能研发人员、人工智能产品生产商、使用人工智能产品或服务的用户。人工智能伦理风险产生的原因与利益相关者息息相关，因此本标准从人工智能利益相关者出发，对数据、研发者、人工智能产品生产商以及使用人工智能产品或服务的用户提出治理要求，并归纳总结人工智能利益相关组织的共性，提出对组织的基本治理要求。

2019 年国家人工智能标准化总体组发布《人工智能伦理风险分析报告》，阐述了多种人工智能伦理风险产生的原因、造成的影响以及应对的方法。2021 年全国信息安全标准化技术委员

会秘书处发布《网络安全标准实践指南—人工智能伦理安全风险防范指引》，针对人工智能伦理安全风险，给出了研究开发者、设计制造者、部署应用者、用户等的安全风险防范措施。2021年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》，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项基本伦理要求。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。

结合上述文献及相关国内外标准文献中对人工智能伦理风险的治理要求及措施，本标准提出了对数据、研发者、生产商、用户的治理要求，以及对人工智能利益相关组织的基本治理要求。

四、与现行相关法律、行政法规和其他标准的关系

本标准遵循法律、法规，符合国家有关现行法律、法规和强制性国家标准的规定，与现行相关法律、法规和国家标准、行业标准相协调，无冲突。

五、重大分歧意见的处理过程、处理意见及其依据

无。

六、对地方标准自发布日期至实施日期之间的过渡期的建议及理由

本标准为推荐性地方标准，建议过渡期为一个月。建议过渡

期间进行本标准的宣贯培训工作，根据本标准的适用范围，面向省人工智能产品研发机构、产品和服务供应商、用户进行标准的培训与宣贯，采用专家讲座、系列课程、交流答疑、发放宣贯材料等方式，积极推进标准实施后的应用。

七、其他需要说明的内容

无。