《人工智能通用大模型合规管理体系 指南》标准编制说明

《人工智能通用大模型合规管理体系 指南》标准起草组

2024 年 12 月 17 日

1、标准范围。

《人工智能通用大模型合规管理 指南》（以下简称本文件）规定了在企业内开展、支持和持续改进人工智能通用大模型合规管理工作的要求，可为以下企业提供参考：利用通用大模型技术向公众提供生成文本、图片、音频、视频等产品或服务，正在开发或委托第三方开发通用大模型，计划提升通用大模型合规管理水平，寻求外部组织对其通用大模型的合规性进行评价。

2、工作简况。

2022 年末，以 ChatGPT 为代表的大规模预训练语言模型引发各界广泛关注，成为新一轮人工智能技术应用爆发的催化剂，并由此带动人工智能通用大模型的算法创新及关键技术研究进入加速期。通用大模型作为人工智能应用发展的核心引擎，凭借其优秀的通用性、泛化性及技术赋能特性，正渐进成为人工智能行业的新型基础设施，为经济发展与产业转型注入新动能。然而，随着通用大模型驱动的人工智能对社会结构、产业生态以及人类生活方式的影响逐步扩大，其在数据安全、可解释性、隐私保护、知识产权归属、内容安全、责任归属等方面的治理风险也在逐渐显露。

通用大模型规范发展是人工智能技术广泛应用的重要前提，从数据、算法、应用等层面加强人工智能通用大模型的合规管理是发展数字经济的应有之义。为促进企业以负责任的方式开发、提供或应用人工智能通用大模型，助力企业提高大模型合规管理水平，标准编写组汇集了产业、监管、法理等领域的专家，深入研究通用大模型合规管

理诉求，紧密跟踪国际和国内标准的进展情况，在2024年4月完成标准大纲架构和草案，并定向征集主要大模型企业的建议，起草组根据建议进一步修订完善了标准，于5月形成了正式的标准草案本文。在此基础上，2024年8月底组织召开17家单位参与的标准草案研讨会，对标准草案内容的先进性与适用性进行了充分讨论与沟通。此后，标准起草组充分吸纳草案研讨会意见建议，并定向与意见建议提出单位进行一对一交流，在全面修订完善草案的基础上，于2024年12月初形成标准征求意见稿。

3、标准编制原则和确定标准主要内容的依据。

标准编制遵循兼顾一般性和特殊性的原则。一方面，本文件规定了在企业内开展、支持和持续改进人工智能通用大模型合规管理工作的要求并提供指导，企业应将其合规管理的重点放在通用大模型的某些特征之上，例如数据质量、隐私保护、模型可解释性和可扩展性等。另一方面，如果通用大模型与传统的人工智能技术相比引发了额外的安全风险，企业可在本文件的基础之上采取不同的保护措施。**标准编制遵循平衡发展与安全的原则**。本文件聚焦科技创新主导的新质生产力，通用人工智能大模型的关键技术及落地应用，以及合规治理，通过法理、监管、司法、产业的多方参与，双轮驱动发展与安全，探讨新时代人工智能通过大模型高质量发展的道、术、法、器、势。

确定标准主要内容的依据是通用大模型相关现行法律、行政法规、标准要求，如《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》、《中华人民共和国科学

技术进步法》、《生成式人工智能服务管理暂行办法》、GB/T 35770-2022《合规管理体系 要求及使用指南》等。

4、主要试验（或验证）的分析、综述报告。

合规管理体系的国际标准的起源是澳大利亚国家标准AS 3806《合规计划》。随着该标准被全球接纳程度越来越高，制定国际性的企业合规管理标准逐渐成为大家的共识和需求。2012年10月，国际标准化组织（ISO）成立ISO/PC271合规管理项目委员会，正式启动合规管理体系的国际标准制定工作。2014年12月，国际标准“ISO19600：2014 Compliance Management Systems Guidelines”正式发布。

2016年5月，我国合规管理体系的国家标准制定正式启动，基本编制原则为等同采用ISO 19600。2017年12月29日正式发布我国的合规管理体系国家标准GB/T 35770-2017 《合规管理体系 指南》。随着经济全球化的发展和企业治理体系的不断演变，全球合规治理领域的内涵和组成也在快速的变化，需要对原有标准进行修订。2020年11月ISO完成修订工作，其后，发布ISO 37301：2021《合规管理体系 要求及使用指南》。我国也于2020年启动相应的国标编制工作，并于2022年10月发布修订后的GB/T 35770-2022《合规管理体系 要求及使用指南》。

在过去的一年多时间里，“大模型”一直是中国科技领域内最热门的赛道，国家相关部门出台了《生成式人工智能服务管理暂行办法》《互联网信息服务算法推荐管理规定》等规范性文件，促进生成式人

工智能健康发展和规范应用。在快速发展与变化的过程中，大模型公司从技术到应用，再到商业化的过程也面临着挑战，行业迫切需要相关协会组织出台指导行业开展合规工作的标准、指南、指引，以便辅导企业更好的符合行业监管和服务社会的要求。标准编写组高度重视行业的诉求，并紧密跟踪国际和国内标准的进展情况，适时启动本团体标准的编写工作。

本文件规定了在企业内开展、支持和持续改进人工智能通用大模型合规管理工作的要求并提供指导，企业应将其合规管理的重点放在通用大模型的某些特征之上，例如数据质量、隐私保护、模型可解释性和可扩展性等。如果通用大模型与传统的人工智能技术相比引发了额外的安全风险，企业可在本文件的基础之上采取不同的保护措施。

5、标准在起草过程中遇到的问题及解决办法：重大分歧意见的处理经过和依据：有无重要技术问题需要说明。

标准起草过程中，起草组主要成员就编制内容的详略程度、合规管理方法，展开了充分的讨论、调研，在综合考虑要求的通用性和具体工作的实操性方面达成了一致。主要讨论经过参见意见汇总表。

暂无重要技术问题需要说明。

6、与国外标准的关系：包括：采用国际标准和国外先进标准的程度，与国外标准主要技术内容的差异（可引用标准前言的内容）。

无国外标准直接采用。ISO 37301：2021 合规管理体系 要求及使用指南仅在文件结构和主要通用要求方面，给本标准提供了借鉴。

7、修订标准时，说明与标准前一版本的重大技术变化，并列出所涉

及的新、旧版本的有关章条（可引用标准前言的内容）：废止/代替现行有关标准的建议。

非修订标准。

8、说明标准与其他标准或文件的关系（可引用标准前言的内容），特别是与有关的现行法律、法规和强制性国家标准的关系。

本标准严格按照中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》、《中华人民共和国科学技术进步法》、《生成式人工智能服务管理暂行办法》相关条款，对大模型企业的合规义务进行梳理，标准的相关内容完全符合现行法律、法规、强制性国家标准。此外，ISO 37301：2021《合规管理体系 要求及使用指南》在文件结构和主要通用要求方面，给本标准提供了借鉴。

9、标准作为强制性标准或推荐性标准的建议：

本标准建议作为推荐性标准使用。

10、贯彻国家标准的要求和措施建议（包括组织措施、技术措施、过渡办法等内容）：标准发布后，对国内外业界可能产生的影响。

本标准主要规定了企业建立、实施、评估、维护及改进通用人工智能大模型合规管理体系的总体要求。标准发布后，将组织标准宣贯和研讨会，组织专家及AI企业参加；同时，通过开展大模型专项合规培训和能力建设，辅导企业按照标准进行合规能力建设；第三，通过企业合规推进计划和中国互联网协会知识产权工委会，组织企业自愿参加标准符合性测评活动，推动企业内部合规管理体系建设。

11、标准是否涉及知识产权的情况说明；如标准中含有自主知识产权，说明产品研发程度、产业化基础及进程。

标准不涉及知识产权。

12、其他应予说明的事项。

无。