TC260

全国网络安全标准化技术委员会技术文件

TC260-005

人工智能应用伦理安全指引 1.0

Ethics-Safety Guidelines for Artificial Intelligence Applications 1.0

2026-05-19 发布

全国网络安全标准化技术委员会发布

摘 要

近年来，人工智能不断加速发展，生成式人工智能、智能体等新技术及其应用层出不穷，人工智能对社会、生产、生活的重构性影响逐步显现。在人工智能积极、深刻地改变社会运行及个人生活方式并带来诸多便利的同时，也带来许多伦理安全挑战。

为进一步确保人工智能安全可控，统筹人工智能发展与安全，保障强化人工智能对国家经济、社会、生态等方面的持续推动作用，帮助人工智能应用相关方在各应用场景开展相关活动时，更好地兼顾发展、安全以及伦理各方面影响，本文件给出了人工智能应用伦理安全的理念与原则，提出了开展相关活动的基本要求，提供了各方实践的应用指引。本文件为原则性、参考性技术文件；涉及个人信息、自动化决策、内容标识、算法治理、知识产权等事项时，应与现行法律法规和部门规章协调适用。

目 次

前　言

本文件由全国网络安全标准化技术委员会（SAC/TC260）发布。

本文件起草单位：清华大学、中国电子技术标准化研究院、上海交通大学、四川大学、北京科技大学、阿里巴巴集团、华为技术有限公司、北京深度求索人工智能基础技术研究有限公司等。

本文件主要起草人：薛澜、梁正、郝春亮、贾开、王姣、赵静、张妍婷、申卫星、朱旭峰、吕飞霄、庞祯敬、王净宇、曾雄、李芒、宋雨鑫、吴宗泽、傅宏宇、林秋诗、吴少卿、王晓箴、邵萌、薛晖、张荣、李娅莉、陈迪思等。

人工智能应用伦理安全指引

1 范围

本文件给出了人工智能应用伦理安全影响，提出了人工智能应用伦理安全理念与原则，规定了人工智能应用开发、服务提供和应用使用等安全指引。

本文件可为组织和个人开展人工智能应用活动提供指导，也可为相关主管部门、行业组织和有关机构推进人工智能伦理安全治理提供参考。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 25069 信息安全技术 术语

GB/T 41867 信息技术 人工智能 术语

3 术语和定义

GB/T 25069和GB/T 41867界定的以及下列术语和定义适用于本文件。

3.1

人工智能应用 artificial intelligence application

具有功能特性的人工智能的使用，该人工智能在利益相关方场景中运行以实现预期结果。

[来源：ISO/IEC 5339:2024(en), 3.1]

3.2

人工智能应用伦理安全（简称“伦理安全”） ethics-safety of artificial intelligence applications

开展人工智能应用活动，处理人工智能与人、社会、自然之间相互关系时，从安全影响角度出发，应保障的伦理价值或秩序的规范与准则。

3.3

应用开发者 developer

开展人工智能应用的理论研究、技术创新、数据归集、模型开发、算法迭代等相关活动，或利用人工智能理论或技术形成具有特定功能、满足特定需求的系统、产品或服务的组织或个人。

3.4

服务提供者 provider

在工作与生活场景中，利用人工智能技术向具体对象提供服务的组织或个人。

3.5

使用者 user

在工作与生活场景中使用人工智能系统、产品或服务的组织或个人。

4 人工智能应用伦理安全影响

主要影响包括：

a）人类主导权影响——人工智能行为超出人类所预设、理解和可控的范围，在对人工智能应用关键节点的监督、干预与纠偏不足时，可能冲击人类在社会运行和治理过程中的主导地位。

b）公共秩序影响——人工智能应用于社会运行和行为决策，在技术快速演化而社会回馈相对缓慢的情况下，可能影响行为规范、市场秩序与社会信任，最终对社会基础公共秩序带来冲击。

c）个体认知与社会价值影响——人工智能普遍应用形成的新社会环境可能对个体认知与社会价值体系带来冲击，产生人类过度依赖、个体与现实社会脱节等问题。

d）社会分化和歧视影响——人工智能应用对不同群体可能产生差异化影响，放大偏见与歧视，造成结构性不利后果，影响公平公正、加剧社会分化。

e）生命健康与基本权益影响——人工智能因其全新的应用方式，可能对自然人的生命健康、人身安全、人格尊严、隐私、财产、劳动等基本权益造成影响，也可能对组织的合法权益造成侵害或产生不利影响，甚至可能对国家安全、公共安全带来危害。

f）可持续生态影响——不合理的人工智能技术路线选择与应用模式可能会带来系统性生态压力，影响人与自然的可持续发展。

5 伦理安全理念与原则

5.1 伦理安全理念

在开展人工智能应用活动时，应将伦理安全要求贯穿全过程，从安全影响角度出发，妥善处理人工智能与人、社会、自然之间的相互关系，以造福人类、服务社会和可持续发展为导向，推动人工智能朝着有益、安全、公平方向健康有序发展，避免不合理的人工智能技术路线选择与盲目应用，保障人工智能始终服务于增进人类共同福祉。

5.2 伦理安全原则

主要原则包括：

a）**增进人类福祉**：坚持正确价值观，坚持智能向善、以人为本，推动人工智能服务人的全面发展和社会共同利益，促进提高生产生活水平、改善公共服务、提升社会治理效能。鼓励人工智能发挥积极作用，拓展人类能力、增进社会福祉，保护个人劳动就业权利，事先评估并防范人工智能应用可能引发的就业替代或失业风险。

b）**尊重生命权利**：坚持生命至上、尊严优先，尊重和保障人的生命健康、人身安全、人格尊严和基本权益。保障人的自主决策、维护个人主体性、避免过度依赖，涉及生命健康、人身安全、人格尊严等重要场景的人工智能应用，应以保障人的生命权利和基本权益为底线，避免对人的身心健康、人身安全和人格尊严造成不利影响。

c）**坚持公平公正**：坚持公平包容和机会均等，避免人工智能应用造成不合理差别对待或加剧既有不公。坚持消除技术偏见和歧视，确保人工智能不对特定民族、信仰、国别、性别等群体以及特定组织或服务造成不公正影响。关注特殊群体与弱势群体权益，鼓励人工智能增进公共服务公平性与可及性。

d）**合理控制风险**：坚持发展和安全并重，统筹人工智能应用的积极价值与潜在影响，强化风险意识和底线思维，推动人工智能在安全、可靠、可预期的范围内健康有序

应用。审慎开展可能对国家安全、公共安全、生命健康等产生重大影响的人工智能应用，提前防范人工智能应用被违法使用、恶意利用、滥用等。

e）**保持公开透明**：提升人工智能应用的透明度，推动其以更加清晰、可感知、可理解、可追溯的方式运行，增进社会对人工智能的认知与信任。

f）**保护隐私安全**：尊重人的隐私权益，避免人工智能应用侵犯合理预期下的隐私边界，保护个人和家庭隐私空间，增强社会公众对人工智能应用的信任。

g）**确保可控可信**：确保人工智能应用的主导权归属人类，鼓励人工智能应用安全开发，在关键环节设置人类控制机制，建立应急处置与人工干预机制，确保人工智能的运行始终处于人类控制之下向善发展，防范人工智能脱离人类监督或威胁人类生存发展。

h）**敏捷共治**：加强安全互认，推动治理规则、技术标准、安全基准等方面国际衔接，提升敏捷治理水平，推动各方协同共治。推动相关主体共同履行人工智能应用伦理安全治理责任，加强协作配合，凝聚治理合力，避免因技术壁垒、封闭排他等加剧数据鸿沟，提升社会人工智能应用素养，共同保障人工智能健康有序发展。

i）**普惠共享**：构建合作共生、能力共建、收益共享的人工智能生态，不利用人工智能强迫他人、实施垄断等。培育开源创新生态，鼓励人工智能模型、工具组件、评测基准等全方位技术开源，同步提升开源生态安全能力。鼓励不同路径技术探索，共享人工智能知识成果、最佳实践与安全治理经验，推动人工智能应用成果更广泛惠及社会。

6 伦理安全指引

6.1 通用指引

主要指引如下：

a）在开展人工智能应用活动前，预先评估应用目的及影响，充分保障国家安全、公共利益、组织及个人权益。

b）以集体主义、人的全面发展等原则为基础，构建符合中国国情和文化传统的伦理规范体系。

c）正确认识人工智能的应用价值、能力边界及潜在影响，避免盲目信赖、片面夸大和过度宣传。

d）保留关键环节的人类判断、监督、干预和纠偏，避免人工智能过度替代人类决策。

e）重视隐私和个人信息保护，防止过度收集、不当使用、泄露或者滥用相关信息。

f）关注人工智能应用可能带来的偏见、歧视和机会不均等影响，重视未成年人、老年人、残障人士等易受影响群体保护。

g）全面提升人工智能应用伦理素养水平，加强人工智能伦理应用影响的安全风险提示、反馈和改进，及时发现问题并采取相应措施。

h）适时开展伦理安全影响评估，及时识别伦理安全风险并完善伦理安全防控措施。

i）涉及生命健康与人身安全的，坚持以人为本、安全优先、审慎适用，以保障生命健康与人身安全为底线，确立人工智能应用的适用边界与责任安排。

j）涉及社会治理与公共服务的，坚持公共利益优先、公平公正，在提质增效的同时，保障公共服务的普惠性、可及性与程序正当，保障公众知情、质疑、复核与救济渠道有效运转。

k）涉及信息资讯与传播的，坚持真实可信与清朗生态导向，推动人工智能提升优质内容供给与传播效率，同时维护传播秩序与受众权益，避免出现违法违规内容。强化对生成与推荐的风险治理，避免人工智能可能带来的信息茧房、认知误导、认知退化等问题。

l）涉及知识发现与生产的，坚持求真务实、可验证与学术诚信，鼓励人工智能在科研探索、工程研发与高等教育等活动中发挥辅助增益作用、更好释放人的创造性与主体性，确保关键结论与重要判断具备必要核验与责任承接，强化证据、引用与署名规范，尊重知识产权与成果归属边界。

m）涉及金融活动的，坚持稳健审慎，注重风险控制与消费者保护。对可能影响个人财产权益、交易机会与交易条件的关键环节，应设置清晰规则，便于责任追溯与结果复核，关注对公平竞争与市场秩序的影响。

6.2 应用开发指引

主要指引如下：

a）统筹考虑人工智能应用伦理安全，识别满足伦理要求的安全开发需求，审慎关注人工智能的合法合规应用情况，避免单纯以性能或效率目标驱动迭代开发与优化改进。

b）满足我国人工智能科技伦理审查要求，审慎开发具备高度自主性人工智能应用，重点评估其失控风险以及其对产业和社会的影响。

c）理性认知人工智能应用的技术能力边界与潜力，正确看待“幻觉”等问题，对缺乏依据或者存在较大不确定性的输出作出必要管控以及提示说明。

d）将安全可控、公平公正、隐私保护、机会均等理念贯穿数据选择、目标设定、算法设计、技术开发、产品研发、测试评估、运行维护等全过程。

e）面向未成年人、老年人、残障人士等易受影响群体的特点，在应用开发过程中采取专门设计考量，其安全提示、透明度标识及隐私授权等采用符合年龄认知的语言，并符合适用的无障碍获取要求，确保易受影响群体能够清晰感知。

f）在技术可行、符合业界普遍技术水平的前提下，提高开发过程透明度，结合人工智能应用场景和影响程度等方面，制定并提供安全配置、安全部署的操作指南，及时、准确、完整对外说明人工智能应用的功能、局限、安全风险和可能的影响。

g）设置事故信息追溯机制，保留与风险事件相关的必要信息，确保在发生争议、损害或事故时能够复盘分析并支撑责任追溯。

h）推动不同组织、个人间合作与互信，促进良性竞争与多元化技术路线并行发展，协同开展伦理安全风险处置。

6.3 服务提供指引

主要指引如下：

a）选择符合伦理影响规范的人工智能应用目标，理性引导社会认知服务能力与边界，及时提示可能风险，避免对使用者造成误导。

b）提供符合科技伦理审查要求的产品或应用，加强人工智能应用全生命周期的跟踪审查。

c）服务涉及国家安全、社会公共权益、组织和个人重大生命财产安全时，人工智能宜仅承担辅助决策作用、不提供直接决策依据。

d）以清楚明确且便于操作的方式向使用者提供拒绝、干预及停止使用人工智能服务的机制。着力避免提供使用者无法便捷地强行通过断电、关停系统等方式停止的人工

智能服务。

e）尽量减少对使用者隐私和敏感个人信息的处理，遵循合法、正当、必要、最小化原则，杜绝偷采、强采、滥采，防止信息泄露以及不当披露。

f）确需收集使用者数据用于训练模型、改进服务的，宜持续对使用者提示收集数据的状态，并显著告知关闭方式。

g）设置事故应急处置机制，持续监控服务运行过程，主动识别发现伦理安全风险，及时处置伦理安全问题，确保合法合规应用，必要时采取人工紧急干预、中止应用等手段。

h）面向未成年人、老年人、残障人士等易受影响群体服务时，着重提升服务便利性，避免造成使用障碍。

i）服务涉及个人权利义务、公共资源分配等场景时，重点提升公平性，并完善风险提示。

6.4 应用使用指引

主要指引如下：

a）正确认知人工智能，学习人工智能伦理安全基本知识，增强对虚假、错误、误导性内容的识别能力，增强对人工智能相关物理安全风险、决策安全风险的认识，形成理性、审慎的使用习惯。

b）适度使用人工智能，正确认知人工智能情感服务原理，以人工智能作为辅助真实生活的工具，避免过度依赖、过度沉迷，避免通过人工智能过多替代现实交往与真实活动。

c）妥善运用人工智能，通过人工智能帮助学习、生活、创作，促进个人主体性提升和创造性发挥。

d）合法使用人工智能，尊重他人尊严与合法权益，不借助人工智能对他人进行误导、骚扰、操纵，不利用人工智能伪造、冒用他人身份或仿冒权威主体，不使用人工智能实施攻击、破坏、盗窃等违法违规行为。

e）妥善保护自身敏感信息，审慎向人工智能提供、上传或分享敏感个人信息、隐私信息，避免个人文件、商业秘密等意外泄露，重点防范涉及国家安全、公共利益的信息被不当传播。

f）提高人工智能素养，积极履行人工智能生成合成内容标识等人工智能相关义务，及时反馈伦理安全风险，帮助建设良性人工智能生态。