TC260-PG-20211A

网络安全标准实践指南

—人工智能伦理安全风险防范指引

(v1.0-202101)

全国信息安全标准化技术委员会秘书处

2021 年 1 月

本文档可从以下网址获得：
www.tc260.org.cn/

全国信息安全标准化技术委员会
NATIONAL INFORMATION SECURITY STANDARDIZATION TECHNICAL COMMITTEE

前　言

《网络安全标准实践指南》（以下简称《实践指南》）是全国信息安全标准化技术委员会（以下简称“信安标委”）秘书处组织制定和发布的标准相关技术实践指南，旨在围绕网络安全法律法规政策、标准、网络安全热点和事件等主题，宣传网络安全相关标准及知识，提供标准化实践指引。

声　明

技术支持单位

本《实践指南》得到中国电子技术标准化研究院、清华大学、中国人民大学、电子科技大学、旷视科技、中科院自动化所、华为、OPPO 等单位的技术支持。

摘　要

近年来，人工智能迅速发展，积极、深刻地改变了个人生活及社会运行，也带来了诸多伦理安全风险，包括影响社会价值、侵犯个人权利、影响公平公正、模糊责任边界等。

为进一步确保人工智能安全可控，统筹协调人工智能发展与安全，促进人工智能对国家经济、社会、生态等方面的持续推动作用，相关组织或个人在开展人工智能研究开发、设计制造、部署应用等相关活动时，应充分识别、防范、管控人工智能伦理安全风险。

本实践指南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。

目　录

1 范围

本文件针对人工智能可能产生的伦理安全风险问题，给出了安全开展人工智能研究开发、设计制造、部署应用等相关活动的规范指引。

本文件适用于相关组织或个人开展人工智能研究开发、设计制造、部署应用等相关活动。

2 术语与定义

2.1 人工智能

利用计算机或其控制的设备，通过感知环境、获取知识、推导演绎等方法，对人类智能的模拟、延伸或扩展。

2.2 研究开发者

开展人工智能理论发展、技术创新、数据归集、算法迭代等相关活动的组织或个人。

2.3 设计制造者

利用人工智能理论或技术开展相关活动，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。

注：系统、产品或服务的形式包括智能算法等虚拟形式，以及智能机器人等实体形式。

2.4 部署应用者

在工作与生活场景中，提供人工智能系统、产品或服务的组织或个人。

2.5 用户

在工作与生活场景中，接受、使用人工智能系统、产品或服务的组织或个人。

2.6 可解释性

人工智能决策或行为的机制机理可以被人类理解的特性。

注1：可以通过提供说明、进行理论论证等方式提高可解释性。

注2：本文件中，不可解释是指部分人工智能具有的，在当前技术发展情况下，人难以理解其全部机制机理的属性。

2.7 可控性

人工智能与预期行为和结果一致的特性。

3 人工智能伦理安全风险

开展人工智能相关活动，应进行伦理安全风险分析，包括：

a. **失控性风险**——人工智能的行为与影响超出研究开发者、设计制造者、部署应用者所预设、理解、可控的范围，对社会价值等方面产生负面影响的风险。

b. **社会性风险**——人工智能使用不合理，包括滥用、误用等，对社会价值等方面产生负面影响的风险。

c. **侵权性风险**——人工智能对人的基本权利，包括人身、隐私、财产等造成侵害或产生负面影响的风险。

d. **歧视性风险**——人工智能对人类特定群体的主观或客观偏见影响公平公正，造成权利侵害或负面影响的风险。

e. **责任性风险**——人工智能相关各方行为失当、责任界定不清，对社会信任、社会价值等方面产生负面影响的风险。

4 人工智能伦理安全风险防范

4.1 基本要求

人工智能伦理安全风险防范的基本要求包括：

a. 应符合我国社会价值观，并遵守国家法律法规；

b. 应以推动经济、社会、生态可持续发展为目标，致力于实现和谐友好、公平公正、包容共享、安全可控的人工智能；

c. 应尊重并保护个人基本权利，包括人身、隐私、财产等权利，特别关注保护弱势群体；

注：弱势群体是指生存状况、就业情况、发声途径或争取合法权益保障能力等方面处于弱势的群体。

d. 应充分认识、全面分析人工智能伦理安全风险，在合理范围内开展相关活动；

注：合理范围是指以保障个人权利、提升社会价值为目标，具备明确边界以及清楚预期的范围。

e. 研究开发者、设计制造者、部署应用者应积极推动人工智能伦理安全风险治理体系与机制建设，实现开放协作、共担责任、敏捷治理；

注：敏捷治理是指持续发现和降低风险、优化管理机制、完善治理体系，并推动治理体系与机制覆盖人工智能系统、产品和服务全生命周期的理念。

f. 研究开发者、设计制造者、部署应用者应积极推动人工智能伦理安全风险以及相关防范措施宣传培训工作。

4.2 研究开发

研究开发者：

a. 不应研究开发以损害人的基本权利为目的的人工智能技术；

b. 应避免研究开发可能被恶意利用进而损害人的基本权利的人工智能技术；

c. 应谨慎开展具有自我复制或自我改进能力的自主性人工智能的研究开发，持续评估可能出现的失控性风险；

注：自主性人工智能指可以感知环境并在没有人为干涉的情况下独立作出决策的人工智能。

d. 应不断提升人工智能的可解释性、可控性；

e. 应对研究开发关键决策进行记录并建立回溯机制，对人工智能伦理安全风险相关事项，进行必要的预警、沟通、回应；

注：研究开发关键决策是指对研究开发结果可能产生重大影响的决策，如数据集的选择、算法的选取等。

f. 应推动研究开发合作、互信，促进良性竞争与多元化技术发展。

4.3 设计制造

设计制造者：

a. 不应设计制造损害公共利益或个人权利的人工智能系统、产品或服务；

b. 应不断提升人工智能系统、产品和服务的可解释性、可控性；

c. 应及时、准确、完整、清晰、无歧义地向部署应用者说明人工智能系统、产品或服务的功能、局限、安全风险和可能的影响；

d. 应在系统、产品或服务中设置事故应急处置机制，包括人工紧急干预机制等；应明确事故处理流程，确保在人工智能伦理安全风险发生时作出及时响应，如停止问题产品生产、召回问题产品等；

e. 应设置事故信息回溯机制；

示例：通过黑匣子实现无人驾驶的事故信息回溯。

f. 应对人工智能伦理安全风险建立必要的保障机制，对引起的损失提供救济。

注：可通过购买保险等手段为必要救济提供保障。

4.4 部署应用

部署应用者：

a. 使用人工智能作为直接决策依据并影响个人权利时，应具有清晰、明确、可查的法律法规等依据；

b. 在公共服务、金融服务、健康卫生、福利教育等领域，进行重要决策时如使用不可解释的人工智能，应仅作为辅助决策手段，不作为直接决策依据；

c. 应向用户及时、准确、完整、清晰、无歧义地说明人工智能相关系统、产品或服务的功能、局限、风险以及可能的影响，并解释相关应用过程及应用结果；

d. 应以清楚明确且便于操作的方式向用户提供拒绝、干预及停止使用人工智能相关系统、产品或服务的机制；在用户拒绝或停止使用后，应尽可能为用户提供非人工智能的替代选择方案；

注：用户停止使用包括因主观原因停止使用，以及因客观条件，如生理缺陷等，无法继续使用的情况。

e. 应设置事故应急处置机制，包括人工紧急干预机制、中止应用机制等，明确事故处理流程，确保在人工智能伦理安全风险发生时作出及时响应；

f. 应向用户提供清楚明确且便于操作的投诉、质疑与反馈机制，并提供包含人工服务在内的响应机制，进行处理和必要补偿；

g. 应主动识别发现人工智能伦理安全风险，并持续改进部署应用过程。

4.5 用户使用

用户：

a. 应以良好目的使用人工智能、充分体现人工智能的积极作用，不应以有损社会价值、个人权利等目的恶意使用人工智能；

b. 应主动了解人工智能伦理安全风险，积极向研究开发者、设计制造者、部署应用者反馈人工智能伦理安全风险相关信息。

参考文献

[1]全国信息安全标准化技术委员会.《人工智能安全标准化白皮书》.2019

[2]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

[3]经合组织（OECD）.《Principles on Artificial Intelligence》. 2019

[4]薛澜.《走向敏捷治理:新兴产业发展与监管模式探究》. 2019

[5]梁正.《人工智能时代亟需构建合理高效的数据治理体系》. 2019

[6]郭锐.《人工智能的伦理和治理》. 2019

[7]贾开.《人工智能与算法治理研究》. 2019