TC260-PG-2020XX

网络安全标准实践指南

—人工智能伦理道德规范指引

（征求意见稿 v1.0-202011）

全国信息安全标准化技术委员会秘书处

2020 年 11 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、人民大学、中科院自动化所、电子科技大学、旷视、华为、OPPO 等单位的技术支持。

摘　要

本实践指南依据法律法规要求以及社会价值观，针对可能产生的人工智能伦理道德问题，给出了安全风险警示、提出了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。

目　录

1 范围

本文件针对可能产生的人工智能伦理道德问题，提出了安全开展人工智能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

利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、技术、系统、产品以及服务。

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。

3 人工智能伦理道德安全风险

进行人工智能相关活动应开展风险分析，可能存在以下风险:

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、可控制的范围，对社会价值产生负面后果的风险。

b. **社会性风险**——人工智能的使用不合理，包括滥用、误用等，影响社会价值观、引发系统性社会问题的风险。

c. **侵权性风险**——人工智能对人的基本权利、人身、隐私、财产等造成侵害或产生负面后果的风险。

d. **歧视性风险**——人工智能对人类特定群体产生主观或客观偏见，造成权利侵害或负面后果的风险。

e. **责任性风险**——人工智能相关各方责任边界不清晰、不合理，导致各方行为失当，对社会信任、社会价值产生负面后果的风险。

4 人工智能伦理道德规范指引

4.1 基本要求

开展相关活动时:

a. 应符合我国社会价值观，并遵守国家法律法规;

b. 应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、人身、隐私、财产等权利，应特别关注对弱势群体的保护；

注：弱势群体指生存状况、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. 应认识到人工智能存在的伦理道德安全风险，进行必要风险分析，在合理范围内开展人工智能相关活动；

e. 研究开发者、设计制造者、部署应用者应推动、参与人工智能伦理道德安全风险治理体系与机制建设，实现共担责任、开放协作、敏捷治理。

4.2 研究开发指引

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. 应避免出现损害人的基本权利、人身、隐私、财产等权利的应用场景，降低人工智能被恶意利用的可能性；

c. 应谨慎开展具备自我复制或自我改进能力的自主性人工智能的研究开发，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，对人工智能伦理道德安全风险相关事项，进行必要的沟通、回应；

注：研究开发决策包括但不限于数据集选择、算法选择等。

f. 应推动与相关方的合作、互信，促进良性竞争与多元化技术发

展。

4.3 设计制造指引

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、可靠性；

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，明确事故处理流程，确保在人工智能伦理道德安全风险发生时作出及时响应；

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. 应对人工智能伦理道德安全风险建立必要的保障机制，对引起的损失提供救济。

示例：通过购买保险等手段为必要救济提供保障。

4.4 部署应用指引

部署应用者：

a. 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；

注：不可解释是指难以对特定决策或行为的产生过程或原因提供说明、证据或论证。

b. 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、风险以及影响，解释相关

应用过程以及应用结果；

c. 应以清楚明确并便于操作的方式向用户提供能够拒绝或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：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。

d. 应设置事故应急处置机制，包括人工紧急干预机制，明确事故处理流程，确保在人工智能伦理道德安全风险发生时作出及时响应；

e. 应向用户提供清楚明确并便于操作的投诉、质疑与反馈机制，并提供包含人工服务在内的响应机制，进行必要的处理和补偿；

h. 应持续监控部署应用过程，主动识别发现人工智能伦理道德安全风险，并持续改进。

4.5 用户使用指引

用户：

a. 应以良好目的使用人工智能、充分体现人工智能的正面价值，不应以有损社会价值、个人权利的目的恶意使用人工智能；

b. 应主动了解人工智能伦理道德安全风险，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道德安全风险相关信息。

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