|  |  |
| --- | --- |
| ICS | 03.100.40 |
| CCS | |  | | --- | |  |   A10 |

团体标准

T/CFIS 2022—XXXX

基于人工智能技术的未成年人互联网应用建设指南

2022 - XX - XX发布

2022 - XX - XX实施

中 国 网 络 社 会 组 织 联 合 会  发布

 版权保护文件

版权所有归属于该标准的发布机构。除非有其他规定，否则未经许可，此发行物及其章节不得以其他形式或任何手段进行复制、再版或使用，包括电子版，影印件，或发布在互联网及内部网络等。使用许可可于发布机构获取。

目次

[前言 III](#_Toc100584579)

[1 范围 1](#_Toc100584580)

[2 规范性引用文件 1](#_Toc100584581)

[3 术语和定义 1](#_Toc100584582)

[4 基本原则 1](#_Toc100584583)

[5 关键参与角色 1](#_Toc100584584)

[5.1 用户（监督者） 1](#_Toc100584585)

[5.2 用户（未成年人） 1](#_Toc100584586)

[5.3 未成年人AI应用服务提供者 2](#_Toc100584587)

[5.4 未成年人AI应用服务评估者 2](#_Toc100584588)

[6 基于人工智能技术的未成年人互联网应用全生命周期关键阶段 2](#_Toc100584589)

[6.1 概述 2](#_Toc100584590)

[6.2 数据获取与准备阶段 2](#_Toc100584591)

[6.3 模型设计与开发阶段 2](#_Toc100584592)

[6.4 测试验证与评估阶段 3](#_Toc100584593)

[6.5 模型部署与使用阶段 3](#_Toc100584594)

[7 基于人工智能技术的未成年人互联网应用建议 3](#_Toc100584595)

[7.1 避免不公平歧视 3](#_Toc100584596)

[7.2 保护未成年人个人隐私和数据安全 3](#_Toc100584597)

[7.3 净化未成年人网络环境 4](#_Toc100584598)

[7.4 关注未成年人身心发展 4](#_Toc100584599)

[参考文献 5](#_Toc100584600)

1. 前言

本文件按照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。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国网络社会组织联合会提出并归口。

本文件起草单位：中国网络社会组织联合会、深圳市腾讯计算机系统有限公司、北京大学、清华大学、中国传媒大学、中国社科院大学、中国科学院自动化研究所、北京工商大学、北京字节跳动科技有限公司、北京快手科技有限公司、上海哔哩哔哩科技有限公司、阿里巴巴（中国）网络技术有限公司、百度在线网络技术(北京)有限公司、北京猿力未来科技有限公司、华为技术有限公司

本文件主要起草人：石竹君、武杨、苏文颖、罗雯、戚蕴、张久珍、李红娟、曾毅、严明、刘晓春、高丽华、孟禹熙、丁秀云、崔毅、何容、钱程、落红卫、王昕、冯志江、吕行、王冬梅、刘颖辰、郭海利、徐谢程、陈岑、陈旭

基于人工智能技术的未成年人互联网应用建设指南

* 1. 范围

本文件给出了建设基于人工智能技术的未成年人互联网应用的总体原则和关键参与角色，以及在全生命周期中涉及的关键阶段，提出了建设基于人工智能技术的未成年人互联网应用的具体措施。

本文件适用于指导互联网企业、软件或系统开发商开发和实施与未成年人互动或可能影响未成年人的基于人工智能技术的互联网应用（以下简称“AI应用”），企业和评估机构创建、审查和/或更新AI应用的政策、策略或行为准则时可参照使用。

* 1. 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。

IEEE Std 2089-2021 IEEE Standard for an Age Appropriate Digital Services Framework Based on the 5Rights Principles for Children

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

* 1. 术语和定义

下列术语和定义适用于本文件。

人工智能 Artificial Intelligence; AI

一门交叉学科，通常视为计算机科学的分支，研究表现出与人类智能（如推理和学习）相关的各种功能的模型和系统或功能单元的能力。

[来源：GB/T 5271.28-2001，28.01.01]

* 1. 总体原则

设计、开发、运维基于AI的未成年人互联网应用，需坚持最有利于未成年人原则，保障未成年人的生存权、发展权、受保护权、参与权等权利，遵循科技伦理，贯彻以下核心理念是至关重要的：

1. 确保未成年人的安全；
2. 尊重未成年人的成长发展规律，助力其健康和福祉；
3. 确保AI技术在应用中的公平、透明、包容、非歧视、可解释性；
4. 确保互联网企业、软件或系统开发商的人工智能治理和能力建设符合科技伦理和未成年人保护要求，促进行业共治。

注1：根据联合国《儿童权利公约》定义，儿童指18岁以下任何人，除非对其适用之法律规定成年年龄低于18岁，而根据国家互联网信息办公室发布的《儿童个人信息网络保护规定》定义，儿童是指不满十四周岁的未成年人。

注2：根据《中华人民共和国未成年人保护法》定义，未成年人是指未满十八周岁的公民，本文以未成年人为主体提出建议。

* 1. 关键参与角色
     1. 用户（监督者）

以监督者角色，监督、指导未成年人使用AI应用的用户，包括但不限于教师、父母、其他法定监护人、对未成年人具有教导或指导角色的用户等。

* + 1. 用户（未成年人）

以直接使用者角色使用AI应用的未成年人用户。

* + 1. AI应用服务提供者

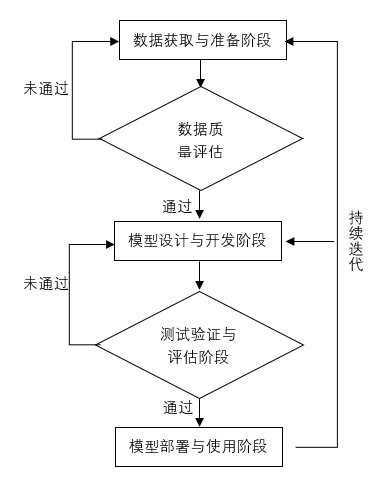
开展未成年人AI应用的设计、开发、运营和维护，专门为未成年人提供AI应用服务或者所提供的AI应用服务的对象包括未成年人的个人或组织。

* + 1. AI应用服务评估者

评估未成年人AI应用是否符合预期目的、相关政策法规和标准要求，评估未成年人AI应用对未成年人的影响的个人或组织，包括：

1. 自建评估组织。AI应用服务提供者所在组织内部组建的评估团队，评估团队应与建设团队相互独立；
2. 第三方评估机构。独立于AI应用服务提供者之外的第三方评估组织。
   1. 基于人工智能技术的未成年人互联网应用生命周期关键阶段
      1. 概述

AI应用对于人工智能技术的使用是以训练模型的建立和训练数据的供给为基础的，因此除了互联网应用全生命周期的经典关键阶段以外，AI应用特有的关键阶段包括数据获取与准备阶段、模型设计与开发阶段、测试验证与评估阶段和模型部署与使用阶段，并且上述关键阶段在AI应用的全生命周期中会持续迭代不断优化。



1. 基于人工智能技术的未成年人互联网应用生命周期关键阶段示意图
   * 1. 数据获取与准备阶段

构建、使用无偏见数据集，对训练、测试、评估、输入等数据和数据集进行数据质量评估，同时监测、发现、识别、纠正数据中的潜在偏见歧视、数据缺口、错误或缺陷、人为操控等风险。

* + 1. 模型设计与开发阶段

基于预期目的、功能用途、行业标准等综合因素制定合适的开发方案，确定合适的技术方法、算法模型、数据集以及风险管理措施，并转换成计算机可运行的程序代码。

* + 1. 测试验证与评估阶段

对应用的所有组成部分，包括数据、预训练的模型、环境和系统整体的行为等，进行测试验证与评估。

提供测试评估服务的未成年人AI应用服务评估者宜由尽可能多元化的人员构成。

* + 1. 模型部署与使用阶段

将通过测试验证与评估的程序代码部署于可供用户访问的环境中，定期监测、分析、评估以及处理不准确、偏见歧视、异常情况等风险，并持续迭代更新应用，优化对未成人的服务。

* 1. 基于人工智能技术的未成年人互联网应用建设措施
     1. 避免不公平歧视

未成年人AI应用可能存在对未成年人的不公平歧视的风险，包括但不限于：

1. 人工智能模型训练所需的数据集存在不具普遍性，模型的泛化能力差，其智能分析和决策存在歧视和偏差；
2. AI技术对未成年人的预测分析和侧写可能存在对未成年人成长和个人发展的影响和限制；

未成年人AI应用服务提供者需考虑以下因素：

1. 使用高质量的数据集，确保数据可用、可靠、可信；
2. 对算法、数据集和设计开发过程进行技术记录，以合适的的方式存储与AI应用可追溯性相关的数据；
3. 为了模型开发之目的，确保没有遗漏所有重要且相关的数据和数据类型，即确保数据能够充分代表所有相关的情况、事项以及目标用户群体；
4. 当使用代理数据进行建模时，确保代理数据与预期目的之间具有最大程度的关联性及合适性；
5. 确保数据所代表的历史情况的客观公正性，避免反映成见偏差等历史偏见；
6. 确保在数据收集过程中不因设备或测量方法导致数据向特定方向偏离而出现测量偏差，并确保收集过程中数据属性、名称等的完整性；
7. 尽可能地使用多方可信来源的异构属性数据集；
8. 为测试程序建立涵盖不同测试目的的多元化的度量基准，构建多元化人员的测试团队；
9. 通过人工干预的方式对智能模型分析和决策的结果进行校正、补充；
10. 通过清晰明确的提示内容，警示应用中AI技术可能对未成年人的影响。
    * 1. 保护未成年人个人隐私和数据安全

未成年人AI应用可能会收集、留存或展示未成年人个人数据或隐私信息，涉及敏感或隐私的数据和信息包括但不限于：

1. 涉及未成年人个人信息的数据，如位置信息、医疗记录、生物识别数据；
2. 涉及未成年人个人生理或心理缺陷等信息的数据，如抑郁症、自闭症相关数据、盲或聋症相关数据；
3. 应用在与用户交互过程中产出或推断出的未成年人的个人信息，如经济社会关系、宗教信仰等数据。

未成年人AI应用服务提供者需考虑以下措施：

1. 具有清晰的未成年人AI应用用户的隐私条款的说明；
2. 采用AI技术识别用户发布的未成年人隐私信息，并及时采取必要的保护措施，如删除信息或向用户屏蔽信息等；
3. 加强数据安全和隐私保护意识教育管理，制定清晰合理的管理机制，对AI应用全生命周期中会触及数据和隐私信息的人员进行有效监管；
4. 明确数据在AI应用全生命周期的权属和使用范围，并通过技术手段保证数据不被滥用或泄露，如对数据访问进行限制，采用防火墙、数据泄露防护系统、数据库加密系统等工具预警、阻断、追溯外部攻击；
5. 尽可能使用最小量的个人数据，采用匿名化数据、同态加密、可信计算环境、合成数据、生成对抗网络、联邦学习、迁移学习等技术方法促进实现数据最小化；
6. 个人信息的处理应征得监护人或监督者角色用户同意；
7. 加强AI应用的人工审核能力，如制定规范化的审核流程和评判标准；
8. 引入监督者角色用户的审核或控制机制，如设置家长模式；
9. 引入专业第三方机构进行风险评估、对未成年人的影响评估、未成年人个人信息保护合规审计，并提供纠正措施。
   * 1. 净化未成年人网络环境

面对AI应用所提供的各类信息和服务，未成年人用户可能存在难以辨别是否适合自身发展的情况，未成年人AI应用服务提供者需考虑以下因素：

1. 通过AI技术，针对未成年人用户，履行防沉迷义务，包括智能提醒使用时长等；
2. 通过AI技术识别未成年人用户越权使用应用的时间管理、权限管理、消费管理、网络直播等功能，比如限制网络游戏应用的非法定游戏时间使用和商品购买等；
3. 通过AI技术识别文字、图像或视频，对可能诱导未成年人用户模仿不安全行为、实施违反社会公德行为、产生不良情绪、养成不良嗜好等可能影响未成年人身心健康的信息标注并予以显著提示；
4. 不推荐展示有关未成年人的淫秽色情网络信息，首页首屏弹窗热搜等处于应用醒目位置、易引起用户关注的重点环节不推荐可能诱导未成年人用户模仿不安全行为、实施违反社会公德行为、产生不良情绪、养成不良嗜好等可能影响未成年人身心健康的信息；
5. 以未成年人用户为服务对象的在线教育AI应用，不应插入网络游戏链接，不应推送广告等与教学无关的信息。
   * 1. 关注未成年人身心发展

不同年龄层的未成年人身心发展具有不同的特点和发展规律， AI应用服务提供者宜发挥自身技术技术，在适合场景下为未成年人提供利于其身心发展的服务：

1. 采用AI技术，针对不同年龄层的未成年人用户的身心发展特点，推荐合适的内容和服务，加强未成年人的网络素养宣传教育；
2. 对目标用户群体不仅限于未成年人的AI应用，宜开设未成年人专区或开启未成年人模式；
3. 充分征集用户、可能直接或间接受应用影响的未成年人的反馈意见，提供易于获取和使用的产品问题或建议的反馈渠道，包括采用AI技术进行问题或建议反馈方式和对AI技术方面的问题，如人工智能客服；
4. 构建用户易于理解和使用的功能模块和业务流程，包括采用语音或文字的关键字等方式调用人工智能的业务，如智能语音识别功能、智能文字和图像识别功能、智能推荐功能等。

参考文献

[1] 中华人民共和国未成年人保护法（2020年10月17日第十三届全国人民代表大会常务委员会第二十二次会议第二次修订）

[2] 儿童权利公约（1991年12月29日第七届全国人民代表大会常务委员会第23次会议批准）

[3] 关于加强网络文化市场未成年人保护工作的意见（2021年11月29日文化和旅游部办公厅发布）

[4] 儿童个人信息网络保护规定（国家互联网信息办公室室务会议审议通过，自2019年10月1日起施行）

[5] 面向儿童的人工智能北京共识 北京智源人工智能研究院，2020年9月14日

[6] 人工智能为儿童——面向儿童群体的人工智能应用调研报告 中国网络社会组织联合会，2021年6月1日

[7] Policy guidance on AI for children Version 2.0 | Recommendations for building AI policies and systems that uphold child rights published by Unicef in November 2021

[8] GB/T 35273-2020 信息安全技术 个人信息安全规范

[9] IEEE Std 2089-2021 IEEE Standard for an Age Appropriate Digital Services Framework Based on the 5Rights Principles for Children

