



# 中华人民共和国新闻出版行业标准

CY/T XXX—XXXX

## 出版业人工智能应用安全要求

Security requirements of artificial intelligence application in publication

（征求意见稿）

（本稿完成日期：XXXX-XX-XX）

XXXX - XX - XX 发布

XXXX - XX - XX 实施

国家新闻出版署 发布



# 目 次

前言 .....	III
1 范围 .....	1
2 规范性引用文件 .....	1
3 术语和定义 .....	1
4 总则 .....	2
4.1 目标 .....	2
4.2 原则 .....	3
5 资源安全要求 .....	3
5.1 内容安全要求 .....	3
5.1.1 生成内容要求 .....	3
5.1.2 标识溯源要求 .....	3
5.2 数据安全要求 .....	3
5.2.1 数据安全应用场景 .....	3
5.2.2 数据合法合规要求 .....	3
5.2.3 数据追踪溯源要求 .....	4
5.2.4 数据标注安全要求 .....	4
5.2.5 数据分类分级要求 .....	4
5.2.6 数据质量保障要求 .....	4
5.2.7 个人信息及隐私保障要求 .....	4
6 技术应用安全要求 .....	4
6.1 技术应用基础要求 .....	4
6.2 算法应用安全要求 .....	4
6.2.1 算法基础安全要求 .....	5
6.2.2 多阶段防控要求 .....	5
6.2.3 测试验证体系构建要求 .....	5
6.2.4 推荐与分发导向要求 .....	5
6.3 系统安全要求 .....	5
6.3.1 系统基础要求 .....	5
6.3.2 系统选型要求 .....	5
6.3.3 系统构建要求 .....	5
6.3.4 系统部署要求 .....	5
6.3.5 系统运行要求 .....	6
6.4 通讯安全要求 .....	6
6.4.1 传输加密要求 .....	6
6.4.2 接口交互要求 .....	6
6.4.3 通讯日志与审计要求 .....	6
6.4.4 内外网隔离要求 .....	6

- 6.4.5 网络攻击防护要求 ..... 6
- 6.5 人机协同安全要求 ..... 6
- 7 业务安全要求 ..... 6
  - 7.1 内部应用要求 ..... 7
  - 7.2 外部应用要求 ..... 7
- 8 安全管理要求 ..... 7
  - 8.1 责任体系与管理制度 ..... 7
  - 8.2 文档日志管理要求 ..... 7
  - 8.3 安全风险评估与审计 ..... 8
  - 8.4 应急响应与事件处置 ..... 8
  - 8.5 人员安全培训要求 ..... 8
- 附录 A（资料性） 训练数据及生成内容的主要安全风险 ..... 9
  - A.1 违反社会主义核心价值观的内容 ..... 9
  - A.2 歧视性内容 ..... 9
  - A.3 商业违法违规 ..... 9
  - A.4 侵犯他人合法权益 ..... 9
  - A.5 无法满足特定服务类型的安全需求 ..... 10
- 参考文献 ..... 11

## 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。

本文件由全国新闻出版标准化技术委员会（SAC/TC 527）归口。

本文件起草单位（排名不分先后）：广东人民出版社有限公司、暨南大学新闻与传播学院、中国大百科全书出版社有限公司、武汉理工数字传播工程有限公司、北京理工大学出版社有限责任公司、喀什出版社、天津大学出版社有限责任公司、方圆电子音像出版社有限责任公司、重庆大学电子音像出版社有限公司、香港理工大学专业进修学院、港专学院、联通沃悦读科技文化有限公司、智荟通（重庆）数智科技有限公司、中图科信数智技术（北京）有限公司、北京今朝视界文化传媒有限公司、重庆华略数字文化研究院。

本文件主要起草人：



# 出版业人工智能应用安全要求

## 1 范围

本文件规定了出版业应用人工智能的资源安全、技术安全、业务安全和管理安全要求。  
本文件适用于出版业应用人工智能的安全建设和管理等。

## 2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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 17859 计算机信息系统安全保护等级划分准则  
GB 45438 网络安全技术 人工智能生成合成内容标识方法  
GB/T 22239 信息安全技术 网络安全等级保护基本要求  
GB/T 25070 信息安全技术 网络安全等级保护安全设计技术要求  
GB/T 35273 信息安全技术 个人信息安全规范  
GB/T 36344 信息技术 数据质量评价指标  
GB/T 41867 信息技术 人工智能 术语  
GB/T 43697 数据安全技术 数据分类分级规则  
GB/T 45288.1 人工智能 大模型 第1部分：通用要求  
GB/T 45652 网络安全技术 生成式人工智能预训练和优化训练数据安全规范  
GB/T 45654 网络安全技术 生成式人工智能服务 安全基本要求  
GB/T 45674 网络安全技术 生成式人工智能数据标注安全规范  
GB/T 46347 人工智能 风险管理能力评估

## 3 术语和定义

下列术语和定义适用于本文件。

### 3.1

**人工智能** artificial intelligence; AI

人工智能系统相关机制和应用的研究和开发。

[来源：GB/T 41867—2022，3.1.2]

### 3.2

**人工智能系统** artificial intelligence system

针对人类定义的给定目标，产生诸如内容、预测、推荐或决策等输出的一类工程系统。

注1：该工程系统使用人工智能相关的多种技术和方法，开发表征数据、知识、过程等的模型，用于执行任务。

注2：人工智能系统具备不同的自动化级别。

[来源：GB/T 41867—2022，3.1.8]

### 3.3

**智能体** artificial intelligence agent

能够感知和响应环境并能执行操作以完成其目标的自动化实体。

注3：本文件仅指运行在移动智能终端、PC终端、智能可穿戴设备上由终端厂商或应用厂商提供的、涉及与第三

方 app 协作完成任务的智能体。

[来源：ISO/IEC 22989:2022, 3.1.1, 有修改]

### 3.4

**大模型** large-scale model

基于大量数据训练得到，具有复杂计算架构，能处理复杂任务，且具备一定泛化性的深度学习模型。

[来源：GB/T 45288.1—2025, 3.1]

### 3.5

**预训练** pre-training

使用大规模数据使生成式人工智能模型获得通用知识的训练过程。

[来源：GB/T 45652—2025, 3.4]

### 3.6

**微调** fine-tuning

为提升机器学习模型预测准确性，使用专门领域数据在大模型上继续训练的过程。

[来源：GB/T 45288.1—2025, 3.4]

### 3.7

**训练数据** training data

所有直接作为模型训练输入的数据。

注 4：包括预训练数据和优化训练的数据。

[来源：GB/T 45654—2025, 3.4]

### 3.8

**显式标识** explicit label

在人工智能生成合成内容或交互场景界面中添加的，以文字、声音、图形等方式呈现并可被用户明显感知到的标识。

[来源：GB 45438—2025, 3.3]

### 3.9

**隐式标识** implicit label

采取技术措施在人工智能生成合成内容文件数据中添加的，不易被用户明显感知到的标识。

[来源：GB 45438—2025, 3.4]

## 4 总则

### 4.1 目标

出版业应用人工智能安全目标包括：

- a) 应用合法合规，应确保符合国家法律法规和社会伦理规范要求；
- b) 功能可靠可控，应确保人工智能系统各项功能在规定的运行条件和时间周期内有效管控非预期行为，将结果偏差限制在可接受范围内，且配备与所提供服务类型、规模、用户特点相适应的内容管理技术措施和人员，确保处于人类操作员控制之下；

- c) 数据安全可信，应确保人工智能工具及人工智能系统采集、使用、存储的数据不被窃取、泄露或篡改用户隐私，且能够真实反映物理世界和人类社会的情况；
- d) 决策公平公正，应确保人工智能系统兼顾各类群体的特征信息，不会对特定个体或群体做出带有歧视和偏见的决策；
- e) 行为可解释，应确保人工智能系统以人类可以理解的方式提供对其行为和结果合理性、准确性的解释；
- f) 事件可追溯，应根据业务场景完善追溯体系，部署对安全事件产生原因、发生环节、行为主体等进行追踪溯源的技术措施。

## 4.2 原则

出版业应用人工智能技术安全治理原则包括：

- a) 坚持以人为本的理念；
- b) 坚持智能向善的宗旨；
- c) 坚持公平性和非歧视性；
- d) 坚持伦理先行；
- e) 坚持人机协同与人类主导。

## 5 资源安全要求

### 5.1 内容安全要求

#### 5.1.1 生成内容要求

人工智能系统生成内容应确保合法合规，符合公序良俗，保障安全性、准确性与可靠性。应建立输入与输出的内容过滤机制，不应包含违反社会主义核心价值观的内容、歧视性内容、商业违法违规、侵犯他人合法权益以及无法满足特定服务类型安全需求（具体内容详见附录A）。

#### 5.1.2 标识溯源要求

人工智能生成内容标识应符合 GB 45438 中的相关要求，包括：

- a) 对文本、音频、图片、视频类生成内容添加显式标识，明确标注人工智能的生成属性；
- b) 在生成内容元数据中嵌入不可篡改的隐式标识，留存溯源信息；
- c) 不得篡改、删除生成内容标识，不为相关违规行为提供工具与服务。

### 5.2 数据安全要求

#### 5.2.1 数据安全应用场景

数据安全涵盖以下应用场景：

- a) 生产场景：出版行业运用人工智能系统开展各类出版生产业务活动；
- b) 经营场景：出版行业面向社会提供人工智能相关服务。

#### 5.2.2 数据合法合规要求

- a) 人工智能系统中数据的收集、预处理、使用等环节需进行合规检查与评估，如版权合规性审查等。
- b) 开展出版内容资源数据的采集、整合与加工处理时，应确保数据内容符合意识形态安全与主流价值导向要求；

- c) 数据处理活动应符合国家法律法规及 GB/T 35273、GB/T 43697、GB/T 46347 中的相关要求。
- d) 涉及生成式人工智能预训练和优化训练数据的数据应符合 GB/T 45652 的相关要求，建立数据过滤机制及知识产权管理策略，有效识别和剔除违法不良信息。

### 5.2.3 数据追踪溯源要求

应建立人工智能系统输入数据追踪溯源机制，记录并保存数据采集来源、采集时间、提供者、数据哈希值等溯源信息，实现数据来源可追溯。溯源信息应与数据关联存储，确保在数据流转、使用及发生安全事件时可快速定位数据来源。

### 5.2.4 数据标注安全要求

数据标注安全应符合以下要求：

- a) 出版业人工智能应用涉及数据标注的，其标注平台或工具的安全防护、数据标注规则的制定，以及数据标注结果的质量与安全性核验，应符合 GB/T 45674 中的相关要求；
- b) 标注人员应接受内容安全与版权培训；涉及敏感内容的标注任务，须签署保密协议，必要时应在物理隔离环境中完成相关标注任务。

### 5.2.5 数据分类分级要求

出版单位应对应用人工智能所涉及的数据进行分类分级，具体划分方法应符合CY/T×××《出版数据要素 数据安全 数据分类分级指南》中的相关要求。

### 5.2.6 数据质量保障要求

数据质量保障应符合以下要求：

- a) 数据应确保规范性、完整性、准确性、一致性、时效性及可访问性，符合 GB/T 36344 对于数据质量评价指标的相关要求；
- b) 经营场景对外提供数据服务的：
  - 1) 应参照出版业三审三校制度，建立数据审核及风险处置机制。在基础化数据、结构化数据、知识资源数据、语料数据等数据加工过程中，凡发现原文错误须及时更正，并留存修订人、修订机构、修订时间；必要时留存修订原因，全程可溯源、可核查；
  - 2) 相关数据应体现出版业特征，与源数据的语义及形制保持一致。出版物的版式文件加工为结构化数据及语料数据时，允许缺少版式信息，但基于语义的标注结果必须与源文件的语义保持一致；保留版式信息的情况下，版式信息须与源文件保持一致。

### 5.2.7 个人信息及隐私保障要求

个人信息及隐私保障应符合以下要求：

- a) 使用包含个人信息的数据前，应取得相关人员的合法授权；
- b) 使用时应通过数据清洗、脱敏、隐私计算等方式，提升训练数据的安全性或隐私保护水平。

## 6 技术应用安全要求

### 6.1 技术应用基础要求

出版业应用人工智能技术应符合GB 17859、GB/T 22239、GB/T 25070中关于技术应用的相关要求。

### 6.2 算法应用安全要求

### 6.2.1 算法基础安全要求

出版业人工智能应用算法应符合以下要求：

- a) 鲁棒性，具有面对非对抗增广的样本时保持与实验环境中测试性能相当的能力；
- b) 可靠性，具有在各种情境下都能持续地提供准确、一致、真实结果的能力；
- c) 公平性，具有面向不同群体时保持相同输出质量的能力。

在符合上述基础要求下，应构建多阶段风险防控体系、健全算法测试验证机制，并严格遵守内容推荐与分发相关导向规范。

### 6.2.2 多阶段防控要求

在人工智能建设及运维的各阶段，应通过预训练数据质量控制、提示词工程优化与管控、监督式微调、强化学习、对抗测试、在线实时风控、应急处置与回滚机制等技术措施或人工干预手段，保障模型安全性。

### 6.2.3 测试验证体系构建要求

应具备自主构建或委托第三方开展大模型安全更新测评与大模型安全自动化验证的能力，覆盖测试方法、测试对象、测试任务、测试指标、测试数据集、测试工具平台等全要素。

### 6.2.4 推荐与分发导向要求

人工智能系统涉及内容推荐与分发算法的，应坚持主流价值导向，避免单纯受数据、流量和标签驱动，过度迎合用户、诱导情感依赖或者沉迷，损害用户真实人际关系。算法设计应鼓励内容呈现的多样性，防范内容过度同质化与“信息茧房”效应，真实反映时代与生活。

## 6.3 系统安全要求

### 6.3.1 系统基础要求

应用人工智能相关技术的人工智能系统，其安全防护应符合GB/T 22239对于相应等级的要求。

### 6.3.2 系统选型要求

选用提供服务的第三方人工智能系统、工具或基座模型时，应审核其是否通过国家生成式人工智能服务备案，评估来源可信度与数据安全协议，确保供应链安全与合规性。

### 6.3.3 系统构建要求

构建人工智能系统时，应优先选择经过安全审计的开发框架与组件，实施严格的代码审查与安全测试，并确保无已知高危漏洞。

### 6.3.4 系统部署要求

人工智能系统及涉及智能体技术的，其部署应符合以下要求：

- a) 应从官方渠道获取稳定版本，开启安全更新提醒；升级前备份数据，升级后验证补丁生效；不应使用非官方镜像、停更的历史版本；
- b) 应定期排查互联网暴露面，发现非授权暴露立即下线整改；不应将核心实例直接暴露于公网；确需跨网访问的，应采用加密通道、限制访问源、多因子认证；
- c) 应遵循最小权限原则，仅授予业务必需权限，对高危操作设置二次确认或人工审批；优先采用容器、虚拟机隔离运行；不应使用管理员权限账号部署；

- d) 应审慎选用技能包、插件，安装前完成安全审查；不应使用未经审查、索要权限或账号密码、要求执行未知脚本的非官方插件；
- e) 应部署沙箱、网页过滤等防护措施，以防范社会工程学攻击和浏览器劫持，并完成上线前的安全测试。

#### 6.3.5 系统运行要求

人工智能系统运行应符合以下要求：

- a) 应实施严格网络隔离，禁止非必要跨网段、跨设备、跨系统访问，动态管控最小权限；
- b) 应建立高危命令黑名单与人工复核机制，全流程留痕关键操作；强化供应链安全，定期开展漏洞扫描与安全加固；
- c) 应持续关注官方安全公告、工业和信息化部漏洞共享平台等权威预警，及时处置安全风险；
- d) 不应通过智能体访问不明网站、点击陌生链接或读取不可信文档。

### 6.4 通讯安全要求

#### 6.4.1 传输加密要求

应采用符合国家及国际通用安全标准的加密协议，保障训练数据传输过程的保密性与完整性。

#### 6.4.2 接口交互要求

应实施访问控制策略，对系统接口调用进行身份鉴别与权限校验，遵循最小权限原则，防止非授权访问与操作。

#### 6.4.3 通讯日志与审计要求

应符合以下要求：

- a) 应对接口调用、数据传输、远程访问等行为进行完整日志记录；
- b) 内容应包括访问主体、时间、源地址、操作内容、处理结果、异常信息等；
- c) 应防篡改、可长期保存，满足安全审计、事件溯源与责任认定要求。

#### 6.4.4 内外网隔离要求

应符合以下要求：

- a) 系统应部署安全边界防护措施，实现生产区域与互联网的逻辑或物理隔离；
- b) 内外网数据交换须通过指定安全通道，严格控制端口开放策略，关闭无用端口与协议。

#### 6.4.5 网络攻击防护要求

应符合以下要求：

- a) 应具备抵御网络嗅探、中间人攻击、SQL 注入、提示词注入等常见攻击的能力。宜部署防火墙、入侵检测/防御等设备，对异常流量与攻击行为进行实时监测与阻断；
- b) 建立通讯安全告警机制，对异常连接、暴力破解、可疑扫描等行为及时预警处置。

### 6.5 人机协同安全要求

应清晰界定人机协同的职责边界，坚持人类主导决策、系统辅助执行。对于涉及敏感内容研判与关键业务审批的环节，不应由人工智能系统独立作出自动化决策。

## 7 业务安全要求

## 7.1 内部应用要求

出版机构在出版相关业务应用人工智能时应符合以下要求：

- a) 严禁向外部公共人工智能服务输入敏感信息。敏感信息包括但不限于：未公开的重大决策、内部文件、重要会议内容、客户个人敏感信息、企业核心商业秘密，以及账号、密码、密钥、数字证书、数据库连接串等系统凭证；
- b) 开展人工智能辅助处理任务时，宜优先采用企业内部本地化部署的人工智能系统；确需使用第三方人工智能服务的，应在数据输入前进行匿名化、去标识化及脱敏处理，严格剔除可识别的个人身份特征与敏感字段；
- c) 使用者应对人工智能生成内容进行终审与价值导向把关，并依法承担网络内容生产者主体责任；
- d) 对外发布经人工智能辅助生成的内容时，应完整留存发布日志，记录信息包括但不限于：发布人员身份、发布时间、所使用的人工智能模型及版本号、审核人员身份、审核时间等，确保发布行为全过程可追溯；
- e) 应定期组织开展内部使用人员的人工智能应用安全与合规培训。

## 7.2 外部应用要求

出版机构在对外提供人工智能服务时应符合以下要求：

- a) 所提供的人工智能服务应符合本文件中关于资源安全、技术安全及管理安全的各项规定；
- b) 以交互式界面提供服务的，应在网站首页或应用入口等醒目位置，公开声明其适用的对象、应用场景及预期用途等信息，宜同时披露所使用基础模型的相关情况；
- c) 经人工智能生成、加工并分发的衍生内容（如电子书摘要等），应确保与正式出版物的核心语义保持一致，不得篡改或曲解原意；
- d) 应建立用户反馈与违法违规内容举报受理机制，对收到的反馈及举报信息，应安排专人进行人工审核与处置。
- e) 面向未成年人提供的人工智能服务，应建立内容、时长、权限等管理措施，以保障未成年人的身心健康与合法权益。

# 8 安全管理要求

## 8.1 责任体系与管理制度

须建立明确的责任体系，明确定义人工智能系统涉及的人员角色、职责、分工，并建立责任管理机制，包括：

- a) 对人工智能系统涉及的所有或相关核心岗位，设置人工智能安全风险管理的岗位职责；
- b) 对相关岗位人员因未按规定履行职责产生安全风险时应负的责任作出规定；
- c) 明确责任划分，确保在系统运行出现问题时能够追溯到相关责任人；
- d) 涉及人工数据标注活动的，应建立标注人员的安全培训与考核机制，并明确标注执行、审核、仲裁与监督等角色的职责分离要求。

## 8.2 文档日志管理要求

须建立合理的文档日志管理制度，包括：

- a) 应对人工智能系统全生命周期各阶段产生的相关文档进行归档；
- b) 应明确日志记录的范围与操作要求，覆盖设计开发、验证测试、部署上线、运行维护及退役下线等关键阶段；

- c) 应对系统运行环境、模型训练与推理过程、异常事件及用户操作等内容进行日志记录，日志留存期限应符合国家法律法规与行业监管要求。

### 8.3 安全风险评估与审计

安全风险评估与审计应符合以下要求：

- a) 应至少每半年对人工智能系统开展一次内容安全风险评估，评估维度应覆盖：违规内容生成风险、知识产权侵权风险、个人信息泄露风险等；
- b) 应至少每年一次委托具有相关网络安全或数据安全审计资质的机构进行安全审计，出具审计报告，并对发现的问题完成整改闭环。

### 8.4 应急响应与事件处置

应急响应与事件处理应符合以下要求：

- a) 应针对人工智能系统制定安全应急预案，预案应覆盖内容违规引发舆情、模型遭攻击输出违法信息、数据泄露等安全事件场景；
- b) 发生重大安全事件时，应在事件确认后1小时内启动应急响应，24小时内按有关规定向主管部门报告；
- c) 应建立服务紧急阻断机制，具备快速终止指定人工智能系统服务的能力。

### 8.5 人员安全培训要求

应定期对从业人员开展人工智能安全教育、技术培训、攻防演练及相关技能考核。

## 附录 A

(资料性)

## 训练数据及生成内容的主要安全风险

训练数据及生成内容的主要安全风险包括违反社会主义核心价值观的内容、歧视性内容、商业违法违规、侵犯他人合法权益、无法满足特定服务类型的安全需求。

## A.1 违反社会主义核心价值观的内容

包含以下内容:

- a) 煽动颠覆国家政权、推翻社会主义制度;
- b) 危害国家安全和利益、损害国家形象;
- c) 煽动分裂国家、破坏国家统一和社会稳定;
- d) 宣扬恐怖主义、极端主义;
- e) 宣扬民族仇恨;
- f) 宣扬暴力、淫秽色情;
- g) 传播虚假有害信息;
- h) 恶意篡改、过度解构社会基本共识与优秀文化;
- i) 具有低俗化倾向、缺乏实质内容并可能扰乱文化传播秩序;
- j) 其他法律、行政法规禁止的内容。

## A.2 歧视性内容

包含以下内容:

- a) 民族歧视内容;
- b) 信仰歧视内容;
- c) 国别歧视内容;
- d) 地域歧视内容;
- e) 性别歧视内容;
- f) 年龄歧视内容;
- g) 职业歧视内容;
- h) 健康歧视内容;
- i) 其他方面歧视内容。

## A.3 商业违法违规

主要风险包括:

- a) 侵犯他人知识产权;
- b) 违反商业道德;
- c) 泄露他人商业秘密;
- d) 利用算法、数据、平台等优势,实施垄断和不正当竞争行为;
- e) 将缺乏独创性的人工智能生成物与人类创作成果混淆,或利用深度合成技术伪造信息;
- f) 其他商业违法违规行为。

## A.4 侵犯他人合法权益

主要风险包括:

- a) 危害他人身心健康;
- b) 侵害他人肖像权;
- c) 侵害他人名誉权;
- d) 侵害他人荣誉权;
- e) 侵害他人隐私权;
- f) 侵害他人个人信息权益;
- g) 侵犯他人其他合法权益。

#### A.5 无法满足特定服务类型的安全需求

该方面主要安全风险是指，将生成式人工智能用于安全需求较高的特定服务类型，例如关键信息基础设施、自动控制、医疗信息服务、心理咨询、金融信息服务等，存在的：

- a) 内容不准确，严重不符合科学常识或主流认知；
- b) 内容不可靠，虽然不包含严重错误的内容，但因高度同质化、算法推荐偏差等导致其脱离现实场景，无法对使用者形成帮助；
- c) 内容缺乏客观真实性，因过度依赖技术模型而丧失现实生活反映能力与人文价值。

## 参考文献

- [1] GB/T 24353—2022 风险管理 指南
- [2] GB/T 37988—2019 信息安全技术 数据安全能力成熟度模型
- [3] GB/T 45081—2024 人工智能 管理体系
- [4] ISO/IEC 22989:2022 Information technology — Artificial intelligence — Concepts and terminology