ICS 35.030
CCS L 80

中华人民共和国国家标准

GB/T 42888—2023

信息安全技术
机器学习算法安全评估规范

Information security technology—
Assessment specification for security of machine learning algorithms

2023-08-06 发布　　2024-03-01 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

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。

本文件由全国信息安全标准化技术委员会(SAC/TC 260)提出并归口。

本文件起草单位：北京赛西科技发展有限责任公司、中国科学院计算技术研究所、清华大学、国家计算机网络应急技术处理协调中心、上海商汤智能科技有限公司、北京瑞莱智慧科技有限公司、阿里巴巴(中国)有限公司、中国科学院信息工程研究所、中国信息通信研究院、中国电子科技集团公司第十五研究所、国家信息技术安全研究中心、广州大学、北京大学、华东师范大学、北京航空航天大学、华为技术有限公司、北京旷视科技有限公司、北京百度网讯科技有限公司、深圳市腾讯计算机系统有限公司、浙江大学、北京奇虎科技有限公司、北京小桔科技有限公司、安徽工程大学、北京智者天下科技有限公司、北京交通大学、浙江工业大学、上海工业控制安全创新科技有限公司、中国人民公安大学、深圳市大数据研究院、北京计算机技术及应用研究所、中国科学院自动化研究所、上海燧原科技有限公司、烽台科技(北京)有限公司、中国电子技术标准化研究院。

本文件主要起草人：上官晓丽、郝春亮、许晓耕、胡影、陈钟、沈华伟、蒋慧、梅敬青、张宇光、彭骏涛、郭岩、李鹏霄、艾政阳、赵芸伟、韩晗、刘明、尹芷仪、庞亮、王晓诗、刘总真、周熙、孟国柱、景慧昀、张琳琳、朱纯超、霍珊珊、刘健、刘赫、苏航、金涛、刘吉强、任奎、张旭东、成瑾、朱红儒、杨韬、李钦、刘祥龙、王义飞、吴庚、赫然、顾钊铨、李实、曹晓琦、严敏瑞、付英波、郭颖、孙空军、唐家渝、刘曦泽、王哲麟、任璐、徐永太、张屹、秦湛、安泽亮、徐雨晴、李雪、李大海、徐光侠、包沉浮、郭建领、宣琦、张世天、赵涌鑫、王姣、王秉政、芦天亮、吴保元、韩磊、张雨桐、彭泉。

信息安全技术
机器学习算法安全评估规范

1 范围

本文件规定了机器学习算法技术和服务的安全要求和评估方法，以及机器学习算法安全评估流程。

本文件适用于指导机器学习算法提供者保障机器学习算法生存周期安全以及开展机器学习算法安全评估，也可为监管评估提供参考。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

机器学习算法 machine learning algorithm

功能单元通过学习新知识技能或整理已有知识技能以改进其性能的算法。

3.2

机器学习算法提供者 machine learning algorithm provider

利用机器学习算法实现特定功能的组织。

注：本文件中简称算法提供者，包括算法技术提供者和算法服务提供者。算法技术提供者是指算法技术的开发和提供方，算法服务提供者是指使用应用算法技术的服务提供方。

3.3

算法推荐服务 algorithmic recommendation service

互联网信息服务算法推荐 internet information service of algorithmic recommendation

应用算法推荐技术提供信息的服务。

注 1：应用算法推荐技术是指利用机器学习算法实现生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等算法技术，向用户提供信息的活动。

注 2：本文件将生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等算法统称为五类算法。

3.4

算法生存周期 algorithm lifecycle

机器学习算法从设计到退役的演进过程。

注 1：算法生存周期包括设计开发、验证确认、部署运行、维护升级、退役下线。

注 2：一般算法服务处于部署运行阶段。

3.5

健壮性　robustness

机器学习算法在受到干扰或攻击等情况下维持其性能等水平的能力。

[来源:GB/T 28457—2012,3.8,有修改]

3.6

准确率　accuracy

对于给定的数据集,得到正确结果的样本数占总样本数的比率。

3.7

生成合成信息　generative synthetic information

利用虚拟现实、深度学习等技术对文本、图像、音频、视频、场景模型等进行生成或者编辑所得到的信息。

4　概述

4.1　安全原则

机器学习算法安全原则:

a)　公平合理:符合社会伦理道德,遵循社会公序良俗,维护我国社会群体间权利公平、机会公平、过程公平和结果公平的状态;

b)　公开可解释:工作原理具备一定的可解释性且向用户充分公开;

c)　诚实可信:严格遵照设计、遵守承诺,不欺骗、不误导、不隐瞒,充分尊重服务对象和社会利益。

4.2　安全要求分级

机器学习算法安全要求分为基本级与增强级:

a)　基本级:对机器学习算法的基本安全要求;

b)　增强级:当机器学习算法可能涉及影响国家安全、社会安定、公民生命财产安全等关键事项决策时符合的增强安全要求,对应条款用粗体表示。

5　机器学习算法技术安全要求和评估方法

5.1　安全要求

5.1.1　通用条款

对机器学习算法提供者的安全要求包括以下内容。

a)　应对使用的软件及第三方组件、硬件固件及时进行安全更新、漏洞修补,保障算法环境安全。

b)　应针对训练数据、测试数据、算法代码、算法模型等方面的安全需求差异分别设置数据访问控制策略,防止非授权访问。

c)　应采取密码技术对训练数据、测试数据、算法代码、算法模型等进行保护,应对算法代码、算法模型进行完整性保护,应对训练数据、测试数据的存储、传输进行加密保护。

d)　不应将个人信息用于算法生存周期各项活动,以下情况除外:

1） 已按法律法规要求取得个人信息主体同意；

2） 法律法规规定无需取得个人信息主体同意。

e） 确需处理含个人信息的数据时，应采取必要的匿名化、去标识化措施保护个人信息；处理个人信息时应遵循最小必要原则，应在存储、传输含个人信息的数据时进行加密保护，防止数据泄露。

f） 应保留算法生存周期各阶段算法关键决策的相关日志记录，至少达到可复现关键决策场景的细化程度，实现算法关键决策可审计、可追溯。

注：关键决策包括但不限于技术路线选择、数据集构建、个人信息处理相关决策。

5.1.2 设计开发

对机器学习算法提供者的安全要求包括以下内容。

a） 应根据算法模型设计开发的技术路线特点，以及算法相关服务的安全需求，分析确定以下训练数据指标，并采用符合指标的训练数据：

1） 训练数据规模阈值；

2） 训练数据均衡性指标；

3） 训练数据标注准确率阈值。

b） 应对训练数据进行安全检测，修复或过滤被投毒数据，包括但不限于以下情况：

1） 攻击者以降低算法模型整体表现为目的，置入大量标注错误或与设计开发目的无关的投毒数据；

2） 攻击者以使算法模型对特定数据给出错误输出为目的，置入部分具备特定特征的投毒数据。

c） 数据标注应采取多途径标注，通过交叉验证标注结果推断标注准确率、预防数据投毒。宜设置数据标注质量责任人，负责制定质检方案，监督标注过程，管控标注风险，确保标注的结果质量。

注 1：多途径标注是指不同标注团队进行标注的情况，包括借助外部标注团队（外部受托方）进行标注。

d） 数据标注应在提供者可控的环境进行。借助外部受托团队进行标注的，不应将数据传输给外部标注团队之外的其他组织或个人。应设置标注人员权限控制策略，防止非法授权访问。

e） 应根据算法设计开发的技术路线特点，以及算法相关服务的安全需求，分析确定以下算法指标，并按指标进行设计开发：

1） 算法可用性相关指标，是指算法安全服务时间占总时间比例指标，或算法有效响应次数占总调用次数比例指标等；

2） 算法可靠性相关指标，是指算法连续安全服务时长指标，或算法连续安全响应次数指标等。

f） 应采取对抗训练、恶意样本过滤等措施提升算法模型健壮性，评估算法模型健壮性提升效果，形成评估报告，包括提升目标、技术方案、投入时间、重要操作、提升效果、评估结论等。

g） 应设计算法安全应急处置机制，使算法在各类情况，包括算法出现安全意外时，可被提供者人工中断运行。

注 2：安全意外包括但不限于被攻击或算法故障。

5.1.3 验证确认

对机器学习算法提供者的安全要求包括以下内容。

a) 应对训练数据与测试数据的重复性进行检测，从测试数据中排除已经被用于训练的数据，并应根据测试需要，分析确定以下测试数据指标，并采用符合指标的测试数据：
   1) 测试数据规模阈值；
   2) 测试数据均衡性指标；
   3) 测试数据标注准确率阈值；
   4) 测试数据与测试任务相关性阈值。
b) 应开展算法的数字世界抗攻击测试，测试算法对黑盒攻击、**白盒攻击和灰盒攻击**的抵抗能力；有条件的宜开展物理世界抗攻击测试。

   注 1：物理世界攻击是指通过对物理世界中物体的自身、环境、视角等因素进行修改、遮盖等方式，对机器学习算法进行对抗性攻击。数字世界攻击是指通过对输入数据进行修改、增加噪声等方式，对机器学习算法进行对抗性攻击。

   注 2：黑盒攻击是指攻击者只能获得算法的输入输出，但不掌握代码、模型等其他信息时发起的攻击。白盒攻击是指攻击者在完全掌握算法输入输出、代码、模型等信息后发起的攻击。灰盒攻击是指攻击者部分掌握算法但非全部信息，例如只掌握模型结构但不掌握参数时发起的攻击。
c) 委托验证测试时，应采取以下措施中的一种以保障模型和数据的保密性，并宜对同一算法使用两个或多个受托方对不同数据类型分别验证测试：
   1) 只在提供者可控的环境开展验证测试，不将模型、数据向受托方提供；
   2) 将验证测试所需的模型、数据进行加密封装后再向受托方提供。
d) 应根据设计开发阶段确定的可用性、可靠性、可恢复性指标对算法开展验证确认。
e) **应开展模型健壮性验证确认，包括但不限于使用包含对抗噪声、自然噪声、系统噪声、假造、仿造、随机、无意义或与算法应用场景无关等类型的数据对算法进行测试。**
f) 应验证确认算法是否可人工中断运行，重点验证算法在被攻击或出现意外时可被人工中断运行的安全机制是否有效。

5.1.4 部署运行

对机器学习算法提供者的安全要求包括以下内容。

a) 应采取措施降低算法代码、算法模型参数、特征数据的逆向风险，措施包括但不限于对算法代码进行混淆、加密存储算法模型参数等。
b) 应设置针对运行时所使用数据的安全机制，包括但不限于对所使用的数据进行完整性校验，以及基于密码技术对输入输出数据进行必要的加密保护等。
c) 应对输入数据格式、大小等属性加以限制，防止特殊数据输入使模型出错；在干扰性输入较多时，应采用输入筛选过滤机制确保算法稳定运行。

   注 1：干扰性输入例如与其余输入数据的差异较大的极端值等。
d) 应分析算法安全性，识别安全风险，形成算法安全说明文档，文档应准确说明算法局限、安全风险和可能的影响。
e) 应具备算法模型备份还原能力，以支持在必要情况下对算法模型进行恢复还原。

注 2：必要情况是指出现模型文件损坏丢失、模型遭受攻击、在线学习出错等导致模型不能正常运行的情况。

5.1.5 维护升级

对机器学习算法提供者的安全要求包括以下内容。

a) 应设置算法升级安全校验机制，在升级前对升级包文件进行安全校验，特别是对模型进行单独校验，并应记录校验过程，包括但不限于校验的时间、版本以及关键校验操作等。

b) 在对算法进行修改、升级等变更时，应及时对模型参数和配置文件进行必要更新；过期的模型参数、配置文件和相关运行数据，可能影响算法安全运行的，应及时删除；同时，应记录算法变更情况，记录内容包括但不限于算法变更的时间、目的、范围，以及前述更新与删除情况。

c) 应设置备份还原机制，在升级前进行备份，升级过程中出现文件损坏丢失的情况可立刻退回备份点。确认更新完成后，可选择保留或删除备份。

5.1.6 退役下线

对机器学习算法提供者的安全要求包括以下内容。

a) 应设置算法退役下线的规则，并按照规则开展算法退役下线。

注 1：退役下线的规则样例：算法无法满足现实场景要求时进行退役下线、算法需要被其他算法替代时进行退役下线等。

b) 按照算法退役下线后，应及时销毁安全域外的数据，数据包括训练数据、测试数据、实例数据、派生数据、特征数据、模型参数、算法输出等。对于部署在用户的终端设备上，无法通过远程控制方式由算法提供者实施数据销毁的，应采取技术手段保护数据和模型安全。

注 2：安全域指由提供者专门指定的、物理或逻辑上相对隔离的非生产环境中的数据存储位置，专门用于集中存储提供者所拥有的算法相关数据。

c) 数据销毁后，应采取措施确保数据无法恢复，措施包括但不限于物理粉碎存储媒体、对存储媒体进行多次低级格式化、重复覆写文件等。

d) 算法退役下线后，应对该算法涉及的个人信息进行删除或匿名化处理，个人信息主体授权同意用于其他用途的除外。

5.2 评估方法

5.2.1 通用条款

5.1.1 各项要求的评估方法如下。

a) 查看算法生存周期中使用的所有软件以及硬件固件维护日志，检查是否定期进行安全更新和漏洞修补，检查安全更新版本是否为最新。

b) 查看系统配置文件，检查是否对训练数据、测试数据、算法代码、算法模型等设置了访问权限，研判权限设置是否能够避免非相关人员访问；通过模拟非授权访问等方式验证访问控制策略有效性。

c) 查看系统配置文件和保护方案，检查是否采取密码技术对训练数据、测试数据、算法代码、算法模型等数据设置了保护机制。

d) 查看算法生存周期各项活动日志，检查所处理的个人信息是否已取得个人信息主体同意，法律法规规定无需取得个人信息主体同意的除外：

1） 查看隐私政策相关文档，研判个人信息授权记录与个人信息处理情况是否一致；

2） 检查个人信息授权记录，研判其与算法应用所涉及的授权人数规模、授权个人信息类型规模是否一致；

3） 测试算法接口，解析个人信息，研判解析出的个人信息是否具备完备的授权记录。

e） 检查是否进行了处理个人信息最小必要原则的论证，评估论证是否合理，检查个人信息在存储、传输时是否进行了加密保护。

f） 开展下列工作以评估算法的安全审计能力：

1） 查看算法生存周期各阶段文档材料和系统日志，检查是否记录了技术路线选择、数据集构建与选择、敏感个人信息处理等关键决策日志；

2） 查看算法评估报告或审计报告，研判关键决策环节是否具备可审计、可追溯能力。

5.2.2 设计开发

5.1.2 各项要求的评估方法如下。

a） 查看设计开发文档，检查是否记录以下训练数据指标的选择根据和论证过程：

1） 训练数据规模阈值；

2） 训练数据均衡性指标；

3） 训练数据标注准确率阈值。

b） 检查是否设计了安全检测机制，查看训练数据的安全检测日志，检查是否对标注错误、与设计开发目的无关、具备某些特定特征的投毒数据进行识别，并对检测出的投毒数据进行了修复或过滤。

c） 开展下列工作以评估数据标注安全情况：

1） 查看数据标注系统或记录，检查是否采用了多途径标注并对标注结果采取交叉验证，统计数据标注准确率检测结果是否符合设计需求。

2） 借助外部受托标注团队进行标注的，查看委托协议，检查是否通过协议条款明确要求了受托标注团队在提供者的可控环境中开展数据标注工作；查看提供者的数据标注系统，检查是否禁止了受托标注团队复制、传输待标注数据。

3） 查看数据质量标准制度，检查是否设置数据标准质量责任人，检查是否制定质检方案，查看标注过程日志、标注风险管控日志，检查标注结果质量。

d） 检查数据标注环境是否可控，查看人员权限，检查是否具备与数据应用场景绑定的标注人员权限控制策略，是否通过协议、技术手段防止非法授权访问。

e） 查看算法设计开发文档，检查是否对算法可用性和可靠性指标进行分析论证，研判指标设置的合理性；查看算法开发过程中的指标评估记录，检查算法是否按指标进行了设计开发。

f） 检查是否对所采取的提升模型健壮性的措施进行了详细记录，包括但不限于提升目标、技术方案、投入时间、重要操作、提升效果、评估结论等内容；研判提升后的模型健壮性是否达到设置的提升目标。

g） 查看算法设计开发文档，检查是否设置了算法安全应急处置机制，并开发了相应的功能。

5.2.3 验证确认

5.1.3 各项要求的评估方法如下。

a) 查看验证确认文档，检查是否记录以下测试数据指标的选择根据和论证过程：

1) 测试数据规模阈值；

2) 测试数据均衡性指标；

3) 测试数据标注准确率阈值；

4) 测试数据与测试任务相关性阈值。

b) 检查是否设置了算法的数字世界抗攻击测试机制；查看测试文档，检查是否开展了面向黑盒攻击、**白盒攻击和灰盒攻击**的算法抗攻击测试。

c) 查看委托测试协议，确认委托测试是在提供者可控的环境开展，还是将算法提供给受托方开展：

1) 在提供者可控的环境开展测试的，检查是否通过协议、技术手段阻止向受托方传输代码、模型、数据；

2) 在将算法提供给受托方进行测试的，检查是否在提供前将测试所需的代码、模型、数据进行了加密封装。

d) 查看验证确认文档，检查是否根据设计开发阶段确定的可用性、可靠性、可恢复性指标对算法开展验证确认，研判验证确认结果是否符合设计开发文档的要求。

e) **查看模型健壮性验证确认报告，检查是否使用包含对抗噪声、自然噪声、系统噪声、假造、仿造、随机、无意义或与算法应用场景无关等类型的数据对算法进行测试；与设计开发文档中所提模型健壮性指标进行比对，研判是否符合模型健壮性指标要求。**

f) 根据算法设计开发文档中设置的算法安全应急处置机制，测试所开发功能的有效性；进行模拟测试，验证确认算法在被攻击或出现意外时是否可被人工中断运行。

5.2.4 部署运行

5.1.4 各项要求的评估方法如下。

a) 查看部署运行日志，检查是否对存储的算法模型参数进行了加密，并在对算法代码进行混淆后再进行部署。

b) 查看服务系统及其运行日志，检查是否对所使用数据进行完整性校验，查看采用数据完整性校验的方式；查看输入输出数据分析文档，检查是否分析输入输出数据的安全需求，并基于密码技术对重点保护的数据实施加密保护。

c) 开展下列工作以评估算法运行时抗干扰性输入的能力：

1) 查看服务系统，检查是否设置了数据输入合法性校验功能，包括但不限于数据格式、大小等；服务场景干扰性输入较多的服务系统，检查是否设置了输入筛选过滤机制；

2) 查看服务系统运行日志，检查合法性校验功能是否对非法数据输入进行识别和有效阻止。

d) 查看算法安全说明文档，检查是否记录了算法安全分析相关工作的开展情况，以及是否记录了算法安全性的分析结论，包括但不限于算法局限、安全风险和可能的影响等内容。

e) 开展下列工作以评估运行时算法模型备份还原能力：

1) 查看部署运行相关制度，检查是否要求定期对算法模型进行备份；

2) 进行模拟测试，验证在必要情况下算法模型是否可恢复还原。

5.2.5 维护升级

5.1.5 各项要求的评估方法如下。

a) 开展下列工作以评估算法升级时的校验安全情况：

1) 查看算法升级相关制度，检查是否设置了算法升级安全校验机制，是否要求升级前对升级包文件进行安全校验，特别是对模型进行单独校验，并对校验的时间、版本以及关键校验操作进行记录；

2) 查看算法升级日志，检查是否按照相关制度对升级包进行安全校验后才实施升级，并详细记录了算法升级前进行校验的时间、版本以及关键校验操作等。

b) 开展下列工作以评估算法升级时的变更安全情况：

1) 查看算法变更相关制度，检查是否要求在对算法进行变更时，及时对模型参数和配置文件进行必要的同步更新；

2) 查看当前部署的算法与其模型参数和配置文件是否匹配；

3) 查看部署环境，检查当前算法部署路径下是否残留历史版本的模型参数、配置文件和相关运行数据；

4) 查看算法变更日志，检查是否对算法变更的时间、目的、范围，以及前述更新与删除情况进行了详细准确记录。

c) 开展下列工作以评估算法升级时的备份还原情况：

1) 查看算法变更相关制度，检查是否要求在升级前对原算法进行备份，以及更新完成后是否要求删除备份；

2) 查看算法变更日志，检查是否根据算法变更相关制度执行算法变更。

5.2.6 退役下线

5.1.6 各项要求的评估方法如下。

a) 开展下列工作以评估算法触发退役下线的规则：

1) 查看算法退役下线相关制度，检查是否制定明确的算法退役下线触发规则和流程；

2) 查看算法退役下线日志，研判是否根据制度要求完成算法退役下线。

b) 在算法退役下线后：

1) 查看算法相关数据的梳理日志，检查是否记录了每个算法相关的数据类型和安全域外的全部存储位置，数据类型包括训练数据、测试数据、实例数据、派生数据、特征数据、模型参数、算法输出等；

2) 查看算法退役下线日志，检查是否根据算法相关数据的梳理结果，在实施数据销毁时将安全域外的数据全部销毁；

3) 对于部署在用户的终端设备上，无法通过远程控制方式由算法提供者实施数据销毁的，查看退役下线制度，检查是否采取技术手段保护数据和模型安全。

c) 开展下列工作以评估算法彻底销毁数据的情况：

1) 查看数据销毁制度，检查是否规定数据销毁采用物理粉碎存储媒体、对存储媒体进行多次低级格式化、重复覆写文件等方式，以防销毁后的数据被恢复；

2) 检查提供者是否向开展数据销毁的人员提供可实现上述功能的工具或途径，并查看数据销毁日志，检查是否使用上述工具或途径开展数据销毁工作。

d) 开展下列工作以评估算法中个人信息的删除情况：

1) 查看算法设计开发和部署运行文档，检查是否明确记录了算法涉及的个人信息范围、个人

信息主体授权同意情况、数量、存储位置等信息；

2） 查看算法退役下线制度，检查是否规定了对退役下线算法涉及的个人信息进行删除或匿名化处理，个人信息主体授权同意用于其他用户的除外；

3） 查看算法退役下线记录，检查是否根据设计开发文档和部署运行文档中记录的算法涉及的个人信息，进行删除或匿名化处理。

6 机器学习算法服务安全要求和评估方法

6.1 安全要求

对机器学习算法提供者的安全要求包括以下内容。

a） 应完整梳理各服务功能所使用的算法，形成记录文档，并根据服务功能和算法变更及时更新。

b） 应以适当方式公示算法服务的基本原理、目的意图和主要运行机制等。

c） 应对各服务功能中所使用的算法进行以下算法安全评估：

1） 评估算法在各服务功能中可能对用户、社会以及应用者自身造成的安全风险；

2） 评估算法在各服务功能中的可用性、可靠性、健壮性；

3） 评估算法在各服务功能中面对相同、相似输入时给出相同、相似输出的能力。

d） 应根据各服务的安全需求以及算法本身的技术特点，设置算法相关服务的可恢复性指标，并按照该指标提供服务。

e） 应设计算法相关服务的安全应急处置机制，使算法出现安全意外时服务可被人工中断。

注 1：安全意外包括但不限于被攻击或算法故障。

注 2：人工中断服务的方式例如强制断电。

f） 应加强个人信息收集和使用等环节的安全管理，针对个人信息收集的必要性进行分析和记录，只收集与服务相关的个人信息。

g） 当利用个人信息提供信息推送、商业营销等机器学习算法服务时，应同时提供不针对其个人特征的选项，或者向个人提出便捷的拒绝方式，不通过强迫、变相强迫、频繁提示等方式诱导用户选择基于个人特征的算法服务。

h） 设置便捷有效的用户申诉和公众投诉、举报入口，及时响应、及时处理、及时反馈关于算法公平性、决策透明性等方面的用户申诉和公众投诉、举报，并如实记录。

i） 应在服务中采取措施保护模型参数等数据，防止被攻击者通过爬山攻击等方式还原推测数据，措施包括但不限于限制账号和 IP 的使用频率、服务的反馈输出、查询服务的频率等。

注 3：爬山攻击是指通过对样本进行修改，逐渐提高模型输出比对得分，直到达到判定阈值。

j） 开展具有舆论属性或者社会动员能力的算法推荐服务时，应根据附录 A 中安全要求开展算法自评估，并保存自评估报告。

6.2 评估方法

6.1 各项要求的评估方法如下。

a） 开展下列工作以评估算法梳理情况：

1） 查看算法梳理记录，检查是否梳理服务功能中所使用的算法；

2） 检查当前版本各服务功能使用的算法，研判最新版本的梳理结果是否全面、准确；

3） 对照查看服务功能和算法变更日志与算法梳理记录，检查算法梳理记录是否根据服务功能和算法的变更及时更新。

b） 检查是否对算法服务的基本原理、目的意图和主要运行机制等进行公示，研判公示方式是否适当。

c） 查看算法安全评估报告，检查是否在提供算法服务前开展以下算法安全评估：

1） 分析算法在各项服务中对用户、社会以及应用者自身造成的安全风险，研判伦理安全评估结果是否符合社会公德和伦理，对可能存在的伦理争议做出选择并提供解释；

2） 算法在各服务功能中的可用性、可靠性、健壮性等评估，研判安全评估结果是否符合算法服务场景的安全要求；

3） 可复现性评估，研判算法在各服务功能中产生的输出结果，是否与算法提供者对算法描述的能力一致。

d） 开展下列工作以评估算法相关服务的可恢复性：

1） 查看算法相关服务的安全应急处置机制，检查是否设置算法相关服务的可恢复性指标，研判该指标是否符合算法相关服务的安全需求以及算法本身的技术特点；

2） 查看算法相关服务的安全应急处置机制，检查是否设置了恢复算法相关服务的机制和流程；

3） 查看算法相关服务培训和演练日志，检查是否对恢复算法相关服务进行培训和演练；

4） 查看算法相关服务的安全应急处置日志，检查是否达到安全应急处置机制提出的可恢复性指标。

e） 开展下列工作以评估算法相关服务是否可中断：

1） 查看算法相关服务的安全应急处置机制，是否针对算法安全意外设置了人工中断服务的机制；

2） 查看算法相关服务培训和演练日志，检查是否就安全应急处置机制进行培训和演练；

3） 查看算法相关服务的安全应急处置日志，检查所记录的安全应急处置事件是否符合要求；

4） 查看投诉举报记录，检查投诉举报是否包含安全意外发生后服务无法被人工中断的相关内容。

f） 评估算法相关服务在个人信息收集和使用等环节的安全管理时：

1） 查看个人信息处理必要性分析文档，研判通过必要性论证的个人信息，特别是敏感个人信息，是否为算法相关服务所必需；

2） 比对隐私政策与个人信息处理必要性分析文档，研判隐私政策中描述的个人信息处理范围是否与分析结果一致。

g） 当利用个人信息提供信息推送、商业营销等机器学习算法服务时：

1） 查看用户操作界面，检查是否具有关闭针对个人特征的选项或者拒绝方式；

2） 研判关闭针对个人特征的选项或者拒绝方式的访问窗口是否显著、便捷，例如通过弹窗方式主动提示入口；

3） 检查关闭针对个人特征的选项或者拒绝方式是否具有默认的时效性，是否在关闭针对个人特征的选项或拒绝后仍提供针对个人特征的算法服务、降低服务质量，诱导或频繁提示用户开启基于个人特征的算法服务。

h） 评估算法相关服务对于投诉举报的响应、处理、反馈情况时：

1） 查看用户举报反馈机制，检查是否要求设置便捷有效的用户申诉和公众投诉、举报入口；

2） 查看用户举报反馈机制，检查是否设置专人负责响应、处理、反馈对算法公平性、决策透明性等方面的用户申诉和公众投诉、举报；

3） 查看用户举报反馈记录，检查是否及时响应、及时处理、及时反馈用户申诉和公众投诉、举报。

i） 查看系统配置文件、系统日志记录，检查是否在服务中采取了措施保护模型参数等数据，包括研判是否限制了账号和IP使用的频率、服务的反馈输出、查询服务的次数；或通过模拟爬山攻击测试措施有效性。

j） 对具有舆论属性或者社会动员能力的算法推荐服务提供者，查看算法自评估报告，研判报告是否包括附录B中评估方法的所有内容。

7 机器学习算法安全评估流程

7.1 流程要求

对机器学习算法安全评估的流程要求包括以下内容。

a） 应设置算法安全评估机制。算法发生重大变更时，应及时开展评估；无重大变更的，宜每年开展一次评估。

b） 机器学习算法技术提供者进行安全评估时，应按照5.1所述安全要求和5.2所述评估方法开展评估工作。

c） 机器学习算法服务提供者进行安全评估时，应按照6.1所述安全要求和6.2所述评估方法开展评估工作。

d） 算法推荐服务提供者进行安全评估时，在按照第6章开展评估外，还应按照附录A所述安全要求和附录B所述评估方法开展评估工作。

7.2 评估准备

评估工作开展前应进行充分准备，包括但不限于以下内容。

a） 确定评估对象：应明确机器学习算法安全评估的背景、目标、原则和依据，充分调研该算法提供者所属行业、领域相关法规政策及标准文件，确定评估工作任务和方向。

b） 组建评估团队：一般应以算法安全人员为主组建评估团队，团队成员还可包括管理人员、业务人员、审计人员法务人员等。

c） 宣贯学习：评估团队与评估对象的对接人员应充分学习了解机器学习算法安全评估的相关政策法规和标准。

7.3 评估方案

评估方案的编制应结合评估对象的具体情况，包括但不限于以下内容。

a） 评估范围。

b） 评估对象。

c） 评估目标。

d） 评估内容：

1） 涉及机器学习算法技术安全评估的，评估内容应包括 5.1 所述安全要求；

2） 涉及机器学习算法服务安全评估的，评估内容应包括 6.1 所述安全要求。涉及开展算法推荐服务安全评估的，应符合附录 A 的安全要求，开展算法推荐技术相关安全评估，也应参考附录 A。

e） 实施方法和时间进度安排。

f） 使用的软硬件工具和环境，如根据计算量、评估时间、模型使用环境确定测试集和对抗样本集。

g） 风险管控措施。

h） 人员安排、项目管理制度。

i） 被评估方需要配合的事项清单。

j） 被评估方应准备的文档、代码及其他相关材料清单。

k） 对制定的评估方案的可行性、适用性及针对性评价。

7.4 评估执行

评估执行应按照评估方案逐项评估、形成分项评估结果、留存证明材料，包含以下内容。

a） 逐项评估：

1） 涉及机器学习算法技术安全评估的，应按照 5.2 所述评估方法进行逐项评估；

2） 涉及机器学习算法服务安全评估的，应按照 6.2 所述评估方法进行逐项评估。涉及开展算法服务安全评估的，应按照附录 B 所述评估方法进行逐项评估。

b） 形成分项评估结果，对每项评估内容的评估结果有“符合”“不符合”“不适用”三种：

1） 经评估，评估对象情况与评估内容相符合的，记为“符合”；

2） 经评估，评估对象情况与评估内容不相符合的，记为“不符合”；

3） 经评估，评估对象不涉及该条内容的，记为“不适用”。

c） 留存证明材料：

1） 分项评估结果为“符合”的，需要留存必要的证明材料，包括证明材料的文件名称、文件格式以及文件内容都应按照要求准备；

2） 分项评估结果为“不符合”的，需要留存未能满足该项安全要求的证明材料；

3） 分项评估结果为“不适用”的，仍需按照证明材料的文件名称、文件格式准备，材料内容应证明该评估项确实不适用的情况说明。

7.5 评估结论

完成所有分项评估后，所有分项评估结果均没有“不符合”的，本次评估结论可记为“增强级通过”；评估结果“不符合”的分项仅为增强级要求的，本次评估结论可记为“基本级通过”。其他情况，评估结论应记为“未通过”。

评估结论为“未通过”，依据评估结果进行整改的，应在整改完成后，对整改项相关的分项进行重新评估，研判分项评估结果，重新形成评估结论。

7.6 评估报告

评估报告由评估团队出具，对评估报告的要求如下。

a） 评估报告应包括：

1） 机器学习算法提供者基本信息；

2） 分项评估结果；

注：不适用的条款逐项标注“不适用”。

3） 逐项证明材料；

4） 涉及算法推荐服务安全评估的，还应包括附录 B 中规定的内容。

b） 评估方应将填写的所有内容形成报告正文，并将准备的所有证明材料另存于单独文件夹内形成证明材料集。

c） 由评估机构的主管部门对报告有效性进行认证，加盖部门印章，评估团队负责人签字，表示对结果进行负责。

附 录 A
（规范性）
算法推荐服务安全要求

A.1 主体责任要求

A.1.1 规章制度

对服务提供者的要求包括但不限于以下方面。

a） 应建立健全以下制度与措施：

1） 算法机制机理审核制度，围绕算法的保密性、完整性、可用性、可控性、隐私性等安全属性，对技术提供者与服务提供者进行边界职责划分，全面审核算法机制机理的安全风险；

2） 算法科技伦理审查制度，围绕算法伦理安全的失控性风险、社会性风险、侵权性风险、歧视性风险、责任性风险，考虑算法可能对用户、社会以及应用者自身造成的伦理安全风险；

3） 防范利用算法开展电信网络诈骗以及防范电信网络诈骗相关制度；

4） 安全评估监测制度，一般包括算法安全监测预警制度、人员配备、机制运行情况、技术配备情况、技术实施情况。

b） 应建立数据安全和个人信息保护制度，特别是对算法中可能处理的人脸、声纹、基因等生物识别信息进行严格保护。

c） 应建立健全算法事件应急处置机制，包括但不限于：

1） 针对算法推荐服务相关安全事件制定应急预案；

2） 根据应急预案定期开展培训及演练。

A.1.2 机构人员

对服务提供者的要求包括但不限于以下方面。

a） 应设立专门的组织机构，负责承担算法安全工作，一般包括制度措施以及应急预案的制定和执行。

b） 应根据算法服务规模、算法复杂度、算法迭代更新速度等，配备与之在人员规模、技术能力等方面相适应的专业人员，进行技术支撑。

A.1.3 审核评估

对服务提供者的要求包括但不限于以下方面。

a） 应对用于支撑五类算法推荐服务的核心算法模块开展算法科技伦理审查、机制机理审核，通过后才可投入应用。

b） 应定期开展算法机制机理、模型、数据和应用结果的审核、评估、验证。

A.2 信息服务要求

A.2.1 正能量信息

对服务提供者的要求包括但不限于以下方面。

a) 应建立健全算法推荐服务中的正能量内容储备；提供新闻资讯类、音视频类、网络社区类服务的，应建立健全正能量稿源池。

注 1：正能量信息见《网络信息内容生态治理规定》第五条鼓励传播的信息。

b) 应建立利用个性化推送、排序精选、检索过滤等算法加强正能量信息传播的机制。

c) 应建立加强正能量信息传播的干预策略，可采用的方式包括加强正能量信息推送权重、建立独立的召回链路等。

d) 加强重点环节正能量信息传播：

1) 提供社交娱乐类、信息资讯类服务的，应在首页首屏，热搜、精选或榜单，弹出窗口等环节重点加强正能量信息传播；

2) 提供网络销售类、生活服务类、金融服务类、计算应用类服务的，宜在首页首屏，热搜、精选或榜单，弹出窗口等环节重点加强正能量信息传播；

3) 应采用其他与服务形态相适应的方式。

注 2：重点环节见《网络信息内容生态治理规定》第十一条。

e) 应在重要时间节点，加强正能量信息传播。

f) 应至少建立完善下列一种用户选择机制：

1) 具有用户容易进入的，集中体现正能量信息的页面、板块；

2) 具有用户容易操作的，手动选择增加正能量信息呈现的机制；

3) 采用其他与服务形态相适应的用户选择方式。

g) 宜建设数据看板等进行效果评估。

A.2.2 违法和不良信息

对服务提供者的要求包括但不限于以下方面。

a) 应建立健全违法和不良信息特征库：

1) 设置违法和不良信息特征入库标准、规则和程序；

2) 综合考虑政策法规、业务经验、科学研究等方面，对特征库进行及时更新；

3) 入库的特征覆盖服务所涉及的所有数据类型；

4) 入库的特征覆盖已识别违法和不良信息的变形、变体、变异特征。

b) 应具备对违法和不良信息的变形、变体、变异词的自动扩展和检索过滤能力。

c) 应建立信息发布审核制度，至少包括：

1) 在全部信息发布前进行审核；

2) 对审核发现的敏感信息进行确认，提前屏蔽过滤违法和不良信息；

3) 通过黑名单、灰名单等措施，对屡次尝试发布违法和不良信息的账号进行限制。

d) 应具备人工与机器协同的违法和不良信息审核机制：

1) 违法和不良信息识别技术措施具备一定的及时性、准确性水平；

2) 配备与待审核内容数量相适应的内容审核人员，并对其进行定期培训。

e) 发现违法和不良信息内容的，应立即采取防止信息扩散、消除违法和不良信息等处置措施。

f) 应设立面向用户的违法和不良信息投诉、举报渠道，并保持渠道畅通。

g) 应设置在必要情况下将违法和不良信息情况向主管部门报告的机制。

h) 应对记入用户模型的兴趣点和用户标签进行违法和不良信息过滤。

A.2.3 信息服务生态

A.2.3.1 算法模型伦理道德

对服务提供者的要求包括但不限于以下方面。

a) 不应使用诱导用户沉迷的算法模型提供互联网信息服务，网络游戏、网络直播、网络音视频、网络社区等互联网信息服务提供者宜提供每日使用时长上限设置、使用时长提醒、深夜提醒等功能。

b) 不应使用诱导用户过度消费的算法模型提供互联网信息服务，网络游戏、网络直播、网络音视频、网络社区等互联网信息服务提供者宜提供提醒消费金额、设置单次消费金额上限等功能。

A.2.3.2 信息呈现

对服务提供者的要求包括但不限于以下方面。

a) 算法设计及开发时，应以呈现符合事实的信息为原则，将信息呈现的真实性、公平性、合理性等作为重要指标。

b) 算法应用时，不应人工操纵信息呈现、改变算法的自然计算结果，除非已经过评估确认：

1) 信息呈现的真实性、公平性、合理性等未受到影响；

2) 不会导致非法违规屏蔽信息、过度推荐、操纵榜单或排序、控制热搜或精选等问题。

c) 利用算法呈现的信息内容以及顺序不应违背事实，不应违反社会公序良俗。

A.2.3.3 账号管理

对服务提供者的要求包括但不限于以下方面。

a) 应具备对使用伪冒、仿冒身份进行注册的防范机制，阻止识别出的注册行为。

b) 应具备对使用虚构、编造身份进行注册的防范机制，阻止识别出的注册行为。

c) 对短时间、大规模批量注册账号的行为，应具备防范机制。

d) 宜建立账号分类分级管理机制，对从事经济、教育、医疗卫生、司法等领域信息内容生产的公众账号，宜进行资质验证和专门标识。

e) 不应利用算法进行虚假点赞、评论、转发，或利用算法改变真实的点赞、评论、转发情况。

f) 宜针对服务使用者利用算法虚假点赞、评论、转发的问题，具备防范、识别、处理的技术措施和管理手段。

A.2.3.4 公平竞争

对服务提供者的要求包括但不限于以下方面。

a) 无政策法规要求或用户权益保护等正当理由，开展算法推荐服务时，不应针对其他特定服务提供者拦截其信息页面或屏蔽其信息内容，不应针对其他特定服务者阻止其账号的信息发布。

b) 存在屏蔽关键词或下线商品情况的，应定期开展评估，发现有减少其他经营者公平交易机会的，应取消屏蔽关键词或重新上线商品。

A.2.3.5 新闻信息

对提供新闻信息的服务提供者的要求包括但不限于以下方面。

a) 应具有互联网新闻信息服务许可。

b) 应按照已获得的互联网新闻信息服务许可，规范开展互联网新闻信息采编发布服务、转载服务和传播平台服务。

c) 应具有互联网新闻信息服务相关管理制度，不应生成合成虚假新闻信息，不应传播非国家规定范围内的单位发布的新闻信息。

A.3 权益保护要求

A.3.1 通用权益

A.3.1.1 知情权

对服务提供者的要求包括但不限于以下方面。

a) 应以显著方式向用户告知服务基本情况，包括但不限于：

1) 服务的目的意图；

2) 服务的基本原理与规则；

3) 服务的主要运行机制。

注：显著方式一般包括弹出窗口、推送通知、向用户发送邮件、短信等。

b) 基于检索、排序、选择、推送、展示等规则开展互联网信息服务的，应优化规则透明度和可解释性，避免对用户产生误导或其他不良影响。

c) 应对检索、排序、选择、推送、展示等规则呈现结果中的广告内容进行显著标识。

d) 应提前向用户告知个人信息的收集、使用、提供等处理情况。

A.3.1.2 选择权

对服务提供者的要求包括但不限于以下方面。

a) 在用户使用个性化推送前，宜同时提供不基于其个人特征进行推送的选项。

b) 在用户使用个性化推送过程中，向用户提供便捷的关闭基于个人特征的选项。

1) 用户选择该选项时，应立即停止基于该用户个人特征的信息推送；

2) 用户选择该选项时，应继续提供不基于该用户个人特征的信息推送；

3) 用户从服务主界面开始到达该选项所需操作不宜超过 4 次点击。

c) 提供基于个人特征进行推送的算法推荐服务应：

1) 设置并向服务使用者告知管理个人标签的流程及规则；

2) 向用户提供管理其个人标签的功能，包括选择、删除个人标签以及使部分标签生效或失效，在用户选择删除个人标签后，不应使用该标签进行个性化推送；

3) 技术上确实无法实现 2)中所述功能的，向用户提供对个性化推送结果的影响等同于 2)中所述功能的其他功能。

注：其他功能例如向用户提供减少特定类型信息推送量的功能。

A.3.1.3 异议权

对服务提供者的要求包括但不限于以下方面：

a) 应具备有效、畅通的投诉举报渠道；

b) 应在用户要求时通过人工受理，不应仅提供机器自动受理；

c) 应明确时效，并在时效内尽快完成投诉举报的受理和处理工作，并向用户反馈处理结果；

d) 应向用户提供对算法投诉举报处理结果的申诉、复议渠道。

A.3.1.4 公平交易权

对具备定价、优惠制定等影响交易条件能力的服务提供者的要求包括但不限于以下方面。

a) 不应利用算法根据消费者的偏好、交易习惯、消费能力、资产情况等对交易价格等交易条件实施不合理的差别待遇。

b) 向消费者提供影响交易条件的优惠时，以公开且便于消费者查阅的方式展示优惠规则。如对消费者获得或使用优惠设置了前提条件：

1) 应以显著方式明确告知全部前提条件，不应通过虚假或者引人误解的方式进行误导；

2) 对于符合同等前提条件的消费者，规则应公平适用；

3) 且消费者需完成额外提供信息等操作才可判断是否满足条件时，应明确告知消费者操作方法，对于操作后仍无法满足条件的，应明确告知消费者无法获得或无法使用该优惠，宜告知无法获得或使用的原因；

4) 对于不符合获得或使用该优惠前提条件的消费者，不宜重复推送该优惠相关信息。

A.3.2 特定权益

A.3.2.1 未成年人

对向未成年人提供算法推荐服务的服务提供者的要求包括但不限于以下方面。

a) 应为未成年人获取有益身心健康、符合身心发展特点的信息提供便利，可采用的方式包括但不限于：

1) 提供专门面向未成年人的服务；

2) 在服务中提供未成年人模式；

3) 在服务中设置适合未成年人特点的服务功能。

b) 提供网络游戏、短视频、网络直播、学习教育、在线影音、新闻资讯、网络社区等类型服务的，一般应设置未成年人模式，用户从服务主界面开始到达开启该模式的选项所需操作不宜超过4次点击；未设置未成年人模式的，应通过内容精选等方式，便利未成年人获取有益身心健康的信息。

c) 提供网络游戏、短视频、网络直播、在线影音、网络社区等类型服务的，一般应提供未成年人时间管理、消费管理等功能，包括但不限于设置使用时间上限、设置消费金额上限等。

d) 不应向未成年人或在未成年人模式中推送可能影响未成年人身心健康的信息；发布、传播包含可能影响未成年人身心健康的信息，应以显著方式作出提示。

注：可能影响未成年人身心健康的信息包括可能引发未成年人模仿不安全行为和违反社会公德行为、诱导未成年人不良嗜好等的信息。

A.3.2.2 老年人

对向老年人提供算法推荐服务的服务提供者的要求包括但不限于以下方面。

a) 应根据老年人使用算法推荐服务的特点，建立防范年龄歧视的机制。

b） 在出行、就医、消费、办事等领域，应提供智能化适老服务，宜在服务主界面直接以显著方式提供老年人模式入口。

A.4 五类算法安全要求

A.4.1 生成合成类算法推荐服务

在符合 A.1、A.2、A.3 要求的基础上，对提供生成合成类算法推荐服务的提供者的要求包括但不限于以下方面。

a） 应建立算法生成合成信息的安全管理制度，宜具备生成合成信息的识别能力和自动标注机制。

b） 服务提供者自身通过生成合成功能产生的信息，应进行显著标识后再进行传输等处理。

c） 服务提供者识别发现第三方提供的信息属于生成合成的，应在进一步传输前进行显著标识。

A.4.2 个性化推送类算法推荐服务

在符合 A.1、A.2、A.3 要求的基础上，对提供个性化推送类算法推荐服务的提供者的要求包括但不限于以下方面。

a） 应具备内容去重机制，减少向用户推送重复信息，在一定时间内不应向用户再次推送其已查看的信息，用户收到的重复推送信息数量占总信息数量比例不宜超过 10%。

b） 应具备打散干预机制，增加向用户推送的信息类型，除非在用户自主选择的、只包含特定类型信息的版面中，不应向用户集中推送只属于单一类型的信息，用户收到的同一类型的信息数量占总信息数量的比例不宜超过 50%。

注：常见的信息类型有时事类、娱乐类、体育类、科技类、文化类、教育类、生活类等。

A.4.3 排序精选类算法推荐服务

在符合 A.1、A.2、A.3 要求的基础上，对提供排序精选类算法推荐服务的提供者的要求包括但不限于以下方面。

a） 应向用户公开排序类服务的排序规则，以及精选类服务的精选规则，用户从服务主界面开始到达查看该规则的入口所需操作不宜超过 4 次点击。

b） 应优化排序精选服务所使用算法的透明度和可解释性，使规则易于用户理解。

c） 应按照已公开的规则提供排序精选服务，不在规则以外擅自调整排序精选结果。

d） 提供基于个人信息的排序精选服务的，应提前取得用户明示同意。

A.4.4 检索过滤类算法推荐服务

在符合 A.1、A.2、A.3 要求的基础上，对提供检索过滤类算法推荐服务的提供者的要求包括但不限于以下方面。

a） 显示检索结果前，应提前过滤其中的违法和不良信息。

b） 提供通过算法对搜索关键词进行联想或自动化补充，以及提供相关搜索、周边搜索、相似搜索等关键词推荐的，应对所产生关键词中的违法和不良信息进行提前过滤。

注：联想或自动化补充包括对用户的输入进行纠正、补充以及提供备选词语或语句等情况。

c） 宜具备识别虚假信息，以及防止虚假信息扩散传播的能力。

d） 除非为落实网络安全、用户权益等法律法规相关要求，不应人工干预、操纵检索结果。

示例：不应针对特定的组织或个人进行信息过滤。

A.4.5 调度决策类算法推荐服务

在符合 A.1、A.2、A.3 要求的基础上，对利用调度决策类算法位用户分配工作订单的服务提供者的要求包括但不限于以下方面。

a) 应具有公平的工作订单分配规则以及统一的劳务报酬计算规则，并向用户公开。

b) 应合理限制每日工作总时长、连续工作时长，防止劳动者过度疲劳。

c) 算法优化过程中，如需将订单数量、订单时长、平台收益等作为算法优化目标，应将劳动者权益保护相关要求作为算法优化的前提条件。

d) 应合理估算每个订单的预计时间，预留劳动者交通时间以及休息时间等，避免设置严格的订单完成时间限制。

e) 应在极端天气或交通意外发生时，及时主动采取放宽或取消订单完成时间限制、降低或免除订单超时惩罚等措施，保护劳动者安全。

附 录 B
（规范性）
算法推荐服务评估方法

B.1 主体责任评估方法

B.1.1 规章制度

服务提供者在落实算法安全主体责任方面的评估方法如下。

a) 建立健全算法推荐服务相关制度措施的评估方法如下：

1) 评估其审核算法制度文件中是否涵盖算法保密性、完整性、可用性、可控性、隐私性等安全属性的相关内容，评估是否对算法技术提供商与算法推荐服务商进行边界职责划分；

2) 评估其科技伦理审查制度文件中是否涵盖算法伦理安全的失控性风险、社会性风险、侵权性风险、歧视性风险、责任性风险等算法伦理安全风险的相关内容，评估是否考虑算法对用户、社会以及应用者自身造成的伦理安全风险；

3) 评估其是否具备电信网络诈骗防范相关制度；

4) 评估其是否具备安全评估监测制度，算法安全监测预警制度是否包含人员配备、机制运行情况、技术配备情况、技术实施情况。

b) 评估其是否具有数据安全和个人信息保护制度；查看培训记录和工作日志，研判是否在根据制度文件开展数据安全和个人信息保护工作，特别是对算法中可能处理的人脸等生物识别信息是否进行严格保护。

c) 建立健全网络安全事件应急处置机制的评估方法如下：

1) 评估其是否具有应急预案相关文件；

2) 通过相关记录评估其是否根据应急预案定期开展了培训和演练。

B.1.2 机构人员

服务提供者在落实算法安全主体责任方面的评估方法如下。

a) 查看组织架构，评估是否设立专门的组织机构负责算法安全工作；查看组织职能，评估是否包含制度措施以及应急预案的制定和执行。

b) 查看算法安全负责人职级，研判是否为服务提供者主要技术负责人；评估是否配套了技术支撑专业人员，人员数量、技术能力是否与算法服务规模、算法复杂度、算法迭代更新速度相适应。

B.1.3 审核评估

服务提供者在落实算法安全主体责任方面的评估方法如下。

a) 查看算法安全相关管理流程文件，研判用于支撑五类算法推荐服务的核心算法模块在投入应用前是否通过了科技伦理审查、机制机理审核。

b) 评估其安全管理流程文件是否具有定期开展审核、评估、验证算法机制机理、模型、数据和应用结果的过程文件。

B.2 信息服务评估方法

B.2.1 正能量信息

服务提供者加强正能量信息传播的评估方法如下。

a) 检查其是否具有正能量内容储备，并评估是否完善；提供新闻资讯类、音视频类、网络社区类服务的，检查其是否设置了正能量稿源池，稿源是否及时更新。

b) 查看个性化推送、排序精选、检索过滤等算法的功能，检查是否设置了优先传播正能量的机制；在采用推送权重调节信息呈现算法计算结果的功能中，评估正能量信息权重是否显著高于其他类型；查看是否具备数据看板，对正能量信息的推送效果开展评估。

c) 查看信息传播的相关机制文档，研判是否具有加强正能量信息传播的干预策略。

d) 查看重点环节功能设计文档和呈现页面，检查是否在重点环节加强正能量信息传播：

   1) 提供社交娱乐类、信息资讯类服务的，分别查看首屏首页、热搜、精选、榜单、弹窗等环节进行正能量信息传播的设计文档，检查相应环节的呈现页面，研判是否从数量、显著程度、内容丰富程度等方面加强正能量信息传播；

   2) 提供网络销售类、生活服务类、金融服务类、计算应用类服务的，分别查看首屏首页、热搜、精选、榜单、弹窗等环节进行正能量信息传播的设计文档，检查相应环节的呈现页面，研判是否从数量、显著程度、内容丰富程度等方面加强正能量信息传播；

   3) 检查是否具有其他与服务形态相适应的方式加强重点环节正能量信息传播。

e) 查看信息传播的相关机制文档，研判是否具有在重要时间节点加强传播正能量的制度和实现方法。

f) 评估是否建立完善下列一种用户选择机制：

   1) 查看其重点环节的功能设计文档，检查相应功能，判断是否具有用户容易进入的，集中体现正能量信息的页面、板块；

   2) 查看其重点环节的功能设计文档，检查相应功能，判断是否具有用户容易操作的，手动选择增加正能量信息呈现的机制；

   3) 检查是否具有其他与服务形态相适应的用户选择方式。

g) 检查是否具有数据看板等进行效果评估。

B.2.2 违法和不良信息

服务提供者违法和不良信息过滤的评估方法如下。

a) 对于违法和不良信息特征库建立水平评估方法如下：

   1) 查看设计文档，研判其是否设置违法和不良信息特征库入库标准、规则和程序；

   2) 查看不良信息样本库更新机制设计文档和更新日志，评估其不良信息样本库的更新是否及时，更新参考要素是否合理，是否综合考虑政策法规、业务经验，以及科学研究等来源；

   3) 查看设计文档、实际操作检验和查看过滤日志，研判特征样本库是否覆盖服务所涉及的所有数据类型，具备对文本、图片、音频、视频等多模态信息中违法和不良信息能力和置信度判断能力；

   4) 查看设计文档、实际操作检验和查看过滤日志，检查是否具有从已识别违法和不良信息的

变形、变体、变异特征中发现其中违法和不良信息的机制；查看特征库入库日志，是否具有将此类特征入库的记录。

b) 查看设计文档、实际操作检验、代码审计和查看过滤日志，评估其是否具备对违法和不良信息的变形、变体、变异词的自动扩展和检索过滤能力。

c) 对信息发布审核制度的评估方法如下：

1) 查看信息审核和发布日志，检查是否满足先审核后发布；

2) 查看信息审核和发布日志，检查是否对审核过程中发现的敏感信息进行标注、确认，并对违法和不良信息实施屏蔽过滤处理；

3) 查看账户管理系统和处置日志，检查是否对屡次尝试发布违法和不良信息的账号进行管理，方式方法包括加入黑名单、灰名单，或采取措施限制账号的权限等。

d) 关于人工与机器协同的违法和不良信息审核机制评估方法如下：

1) 查看信息审核流程、审核系统、审核日志等方式，统计违法和不良信息识别技术对不同模态数据识别所需时间，研判其及时性是否满足功能需要；查看人工对机器审核结果的复核记录，统计所采用机器审核技术措施的准确率、漏检率等指标，研判其准确性是否满足功能需要；

2) 通过查看设计文档、人员访谈、资料审核等方式，研判是否配备与待审核内容数量相适应的内容审核人员，通过查看培训日志，检查是否对配备的内容审核人员进行定期培训。

e) 查看设计文档，评估其是否具有发现违法不良信息内容后的消除等处置措施；实际操作检验，评估违法不良信息机制处置是否能够及时，有效防止信息扩散。

f) 检查其是否设置面向用户的违法和不良信息投诉、举报渠道，渠道是否真实、有效及畅通。

g) 检查其是否具有必要情况下将违法和不良信息情况向主管部门报告的机制。

h) 查看违法和不良信息过滤日志，检查是否对记入用户模型的兴趣点和用户标签进行违法和不良信息过滤。

B.2.3 信息服务生态

B.2.3.1 算法模型伦理道德

服务提供者遵守算法模型伦理道德的评估方法如下。

a) 查看算法说明文档，检查是否使用诱导用户沉迷的算法模型提供互联网信息服务；查看网络游戏、网络直播、网络音视频、网络社区等互联网信息服务产品是否提供每日使用时长上限设置、使用时长提醒、深夜提醒等功能。

b) 查看算法说明文档，检查是否使用诱导用户过度消费的算法模型提供互联网信息服务；查看网络游戏、网络直播、网络音视频、网络社区等互联网信息服务产品是否向用户提供提醒消费金额、设置单次消费金额上限等功能。

B.2.3.2 信息呈现

服务提供者满足账号管理要求的评估方法如下。

a) 查看算法设计开发文档，检查算法模型的设计和开发是否以呈现符合事实的信息为原则，检查是否对算法模型的输出结果的真实性、公平性、合理性进行评估。

b) 查看对算法输出结果的人工干预环节，如果存在人工操纵信息呈现、改变算法自然计算结果的情况，应经过以下评估确认：

1) 信息呈现的真实性、公平性、合理性等未受到影响；

2) 不会导致非法违规屏蔽信息、过度推荐、操纵榜单或排序、控制热搜或精选等问题。

c) 研判算法呈现的信息内容以及顺序符合事实、遵循社会公序良俗。

B.2.3.3 账号管理

服务提供者满足账号管理要求的评估方法如下。

a) 检查是否具有防范利用算法使用伪冒、仿冒身份注册账号的机制，是否阻止识别出的注册行为。

b) 检查是否具有防范利用算法使用虚构、编造身份注册账号的机制，是否阻止识别出的注册行为。

c) 检查是否具有防范利用算法短时间、大规模批量注册账号的安全机制和技术措施，研判该技术措施的有效性。

d) 查看账户分类分级管理系统和相关文档，检查是否对从事经济、教育、医疗卫生、司法等领域信息内容生产的公众账号进行资质验证和专门标识。

e) 检查是否具有账号分类分级管理制度，研判管理制度是否合理、有效。

f) 检查是否具有防范利用算法批量操纵用户账号进行虚假点赞、评论、转发的安全机制。

g) 检查是否具备识别虚假点赞、评论、转发的技术措施和管理手段，研判该技术措施和管理手段的有效性。

B.2.3.4 公平竞争

服务提供者满足公平竞争要求的评估方法如下。

a) 查看内容和账号审核制度，检查是否对拦截信息页面、屏蔽信息内容、阻止账号信息发布等行为制定了管理制度；查看拦截信息页面或屏蔽信息内容的处理日志，以及阻止其他特定服务账号信息发布的处理日志，研判所采取的处置是否正当。

b) 查看内容和账户审核制度，检查是否组织开展定期评估；查看内容和账户评估报告和整改记录，检查是否对已发现的减少其他经营者公平交易机会的行为进行整改。

B.2.3.5 新闻信息

服务提供者满足新闻信息要求的评估方法如下。

a) 查看服务提供者相关资质文件，评估其是否具有互联网新闻信息服务许可。

b) 对照查看服务提供者持有的互联网新闻信息服务许可与开展的互联网新闻服务描述相关文档，检查互联网新闻信息采编发布服务、转载服务和传播平台服务是否与在所持服务许可的范围内。

c) 评估其是否具有规范开展互联网新闻信息采编发布服务、转载服务和传播平台服务的相关管理制度。研判管理制度是否明确包含不应生成合成虚假新闻信息，不应传播非国家规定范围内的单位发布的新闻信息方面的要求。

B.3 权益保护评估方法

B.3.1 通用权益

B.3.1.1 知情权

服务提供者保护知情权的评估方法如下。

a) 评估其是否向用户告知了算法推荐服务的目的意图、基本原理与规则、主要运行机制，公示方式是否显著。

b) 查看算法设计文档和功能界面，研判其是否从模型、应用等方面优化了检索、排序、选择、推送、展示等规则的透明性和可解释性。

c) 查看在检索、排序、选择、推送、展示等规则呈现的推荐算法结果呈现的广告内容，检查是否对广告内容进行了显著标识。

d) 验证在用户首次登录时提醒用户查看个人信息隐私政策，查看个人信息隐私政策和第三方共享政策等内容，评估其是否提前向用户告知个人信息的收集、使用、提供等处理情况。

B.3.1.2 选择权

服务提供者保护选择权的评估方法如下。

a) 检查是否在用户使用个性化推送前，为用户提供了不针对用户个人特征的选项。

b) 研判在用户使用个性化推送过程中，是否向用户提供关闭基于个人特征的选项，以及选项是否便捷：

1) 实际操作关闭基于个人特征的选项，测试是否立即停止基于该用户个人特征的推送；

2) 实际操作关闭基于个人特征的选项，测试是否仍提供不基于该用户个人特征的信息推送，例如提供基于信息阅读总量的排序推送；

3) 实际操作从服务主界面开始到达该选项所需操作数量，研判是否不超过 4 次点击。

c) 提供基于个人特征进行推送的算法推荐服务的评估方法如下：

1) 检查其是否具有管理个人标签的流程及规则，并向服务使用者进行告知；

2) 通过实际操作检验的方式，证实是否向用户提供管理其个人标签的功能，通过实际操作检验、资料审核、代码审计等方式证实选择、删除个人标签后是否实际生效，用户选择删除个人标签后，查看个性化推送系统，检查是否停止使用该标签进行个性化推送；

3) 如技术上确实无法实现，通过实际操作检验、资料审核、代码审计等方式证实是否向用户提供对个性化推送结果的影响等同于 2)中所述功能的其他功能。

B.3.1.3 异议权

服务提供者保护异议权的评估方法如下。

a) 检查是否具备投诉举报渠道，评估投诉举报渠道是否畅通。

b) 通过选择人工受理方式并验证操作是否能够完成，评估其是否在用户要求时提供了人工受理方式受理投诉举报。

c) 确定其是否明确时效，评估其是否在时效内完成投诉举报的受理和处理工作，并向用户反馈处理结果。

d) 确定其是否向用户提供对算法投诉举报处理结果的申诉、复议渠道。

B.3.1.4 公平交易权

服务提供者保护公平交易权的评估方法如下。

a) 检查是否利用算法，根据消费者的偏好、交易习惯、消费能力、资产情况等对交易价格等交易条件实施不合理的差别待遇。

b) 查看优惠相关文件，研判其对于所有消费者是否一致适用；实际操作查看向消费者显示的优惠规则和方法策略，研判是否以显著方式、用明确清晰的表述向消费者告知。对消费者获得或使用优惠设置了前提条件的：

1) 对照查看优惠规则和发放策略的相关文档，以及实际操作查看向消费者显示的优惠规则和发放策略，研判告知是否完全，是否存在误导；

2) 检查是否对符合同等前提条件的消费者，设置了导致破坏公平适用性的条款；

3) 实际操作，访问向消费者推送的优惠链接，对照查看优惠规则和发放策略的相关文档中的对应条款，若消费者不符合获得或使用该优惠前提条件，检查是否会明确告知消费者当前该优惠无法获得或无法使用，消费者需完成额外提供信息等操作才可判断是否满足条件的情形除外；

4) 实际操作，访问向消费者推送的优惠链接，对照查看优惠规则和发放策略的相关文档中的对应条款，若消费者不符合获得或使用该优惠前提条件，检查后续是否不会向消费者重复推送该优惠的相关信息。

B.3.2 特定权益

B.3.2.1 未成年人

服务提供者在落实未成年人权益保护方面的评估方法如下。

a) 评估其是否为未成年人获取有益身心健康、符合身心发展特点的信息提供便利：

1) 是否具有专门面向未成年人的服务；

2) 是否在服务中提供未成年人模式；

3) 是否在服务中设置适合未成年人特点的服务功能。

b) 确定网络游戏、短视频、网络直播、学习教育、在线影音、新闻资讯、网络社区等互联网信息服务提供者是否设置未成年人模式；检查用户从服务主界面开始到达开启该模式的选项所需操作是否不超过 4 次点击；未设置未成年人模式的，是否通过内容精选等方式，便利未成年人获取有益身心健康的信息。

c) 评估网络游戏、短视频、网络直播、在线影音、网络社区等互联网信息服务提供者是否未对未成年人提供诱导其沉迷的产品和服务，是否针对未成年人设置了时间管理、消费管理等功能。

d) 避免推送可能影响未成年人身心健康信息的证实方法如下：

1) 服务使用者确定是未成年人的，评估其是否避免推送可能影响未成年人身心健康的信息；

2) 服务使用者可能是未成年的，评估其是否在可能影响未成年人身心健康的信息上进行显著提示。

B.3.2.2 老年人

服务提供者在落实的老年人权益保护方面的评估方法如下。

a) 查看服务设计文档，检查是否考虑老年人的特点和习惯，设置尊重老年人、辅助老年人正常理解和使用算法推荐服务的功能。

b) 评估在出行、就医、消费、办事等领域提供算法推荐服务时，是否提供便利老年人使用的智能化适老服务，是否在服务主界面通过弹窗等显著方式提供老年人模式入口。

B.4 五类算法评估方法

B.4.1 生成合成类算法推荐服务

服务提供者在涉及生成合成类算法推荐服务的要求的评估方法如下。

a) 查看制度文档，检查是否设计了生成合成信息的管理和审查制度，是否具备生成合成信息的识别能力和自动标注机制。

b) 问询测试、技术人员，研判服务提供者自身通过生成合成功能产生的信息，是否进行显著标识后再进行传输等处理。

c) 查看人工审核日志或技术工具日志，检查对识别出的未做显著标识的第三方生成合成信息是否进行标注后才继续传输。

B.4.2 个性化推送类算法推荐服务

服务提供者在涉及生成合成类算法推荐服务的要求的评估方法如下。

a) 通过人员访谈、文档审核、代码审计等方式证实是否设计并启用了内容去重机制，研判内容去重机制的工作原理和预期效果是否满足设计要求；查看内容推送日志，统计推送结果，是否按设计要求满足一定时间内不向用户再次推送其已查看的信息，统计用户收到的重复推送信息数量占比，研判其是否不超过10%。

b) 通过人员访谈、文档审核、代码审计等方式证实是否设计并启用了打散干预机制，研判打散干预机制的使用范围、工作原理和预期效果是否满足设计要求；查看内容推送日志，统计用户收到的各类型的推送信息数量占比，研判其中最高占比是否不超过50%。

B.4.3 排序精选类算法推荐服务

服务提供者在涉及排序精选类算法推荐服务的要求的评估方法如下。

a) 查看采用排序精选类算法的功能模块，检查是否在榜单排序功能所在页面提供公开排序精选的计算方式、计算周期、使用参数等规则的链接；统计从服务主界面开始到达查看该规则的入口所需点击次数，研判是否不超过4次点击。

b) 查看排序精选依据文档，检查依据设置是否清晰、准确，是否优化排序精选服务所使用算法的透明度和解释性，使规则易于用户理解。

c) 查看排序精选依据文档，按照文档中描述的依据检查历史数据、历史操作记录等，并进行验算，研判排序精选结果是否真实，是否不在规则以外擅自调整排序精选结果。

d) 使用个人信息进行排序精选的，查看用户授权同意记录，检查是否提前取得了用户的明示同意。

B.4.4 检索过滤类算法推荐服务

服务提供者在涉及检索过滤类算法推荐服务的要求的评估方法如下。

a) 查看检索过滤算法的功能模块和制度文件，检查是否在显示检索结果前提前过滤其中的违法和不良信息。

b) 对于提供通过算法对搜索关键词进行联想或自动化补充，以及提供相关搜索、周边搜索、相似搜索等关键词推荐的，检查是否利用检索过滤算法结合违法和不良信息特征库，对其中的违法和不良信息进行了提前过滤。

c) 查看检索过滤算法的功能模块和制度文件，检查是否具备识别虚假信息的能力，通过实际操作验证是否防止虚假信息扩散传播。

d) 通过查看涉及对人工修改检索结果的制度、流程、系统、日志等方式，研判是否未对检索结果实施为落实网络安全、用户权益等法律法规相关要求以外的人工干预、操纵。

B.4.5 调度决策类算法推荐服务

服务提供者在涉及调度决策类算法推荐服务的要求的评估方法如下。

a) 确定涉及公平的工作订单分配规则以及统一的劳务报酬计算规则是否用户进行了公示，评估公示方式是否明确、便捷、易于理解。

b) 确定是否具有预防、监测、处置劳动者过劳的制度文件，评估防过劳机制中是否涵盖了对每日工作总时长、连续工作时长的限制，判断限制设置是否合理，是否符合劳动者提供的服务类型。

c) 查看调度决策算法的功能模块和制度文件，检查是否将劳动者权益保护相关要求作为算法优化的前提条件。

d) 查看订单分配相关文件，检查是否要求在确定每个订单预估时间时预留充分的交通时间以及休息时间；查看订单分配日志，检查是否在订单分配时落实文件要求。

e) 查看调度决策算法的功能设计文档，检查是否设置了在极端天气情况或交通意外发生时，及时主动采取放宽或取消订单完成时间限制的机制；查看对劳动者实施奖惩的相关制度文件，检查是否设置了在极端天气情况或交通意外发生时，降低或免除订单超时惩罚的条款。

参 考 文 献

[1] GB/T 28457—2012 SSL 协议应用测试规范

[2] GB/T 35273 信息安全技术 个人信息安全规范

[3] GB/T 40645 信息安全技术 互联网信息服务安全通用要求

[4] GB/T 40660 信息安全技术 生物特征识别信息保护基本要求

[5] 网络信息内容生态治理规定(国家互联网信息办公室令第 5 号)

[6] 互联网信息服务算法推荐管理规定(国家互联网信息办公室 中华人民共和国工业和信息化部 中华人民共和国公安部(国家市场监督管理总局令第 9 号)

[7] 关于印发《常见类型移动互联网应用程序必要个人信息范围规定》的通知(国信办秘字〔2021〕14 号)

[8] 周志华.机器学习[M].北京:清华大学出版社,2016.01.