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

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　概述 2
4.1　安全原则 2
4.2　安全要求分级 2
5　机器学习算法技术安全要求和评估方法 2
5.1　安全要求 2
5.2　评估方法 5
6　机器学习算法服务安全要求和评估方法 9
6.1　安全要求 9
6.2　评估方法 9
7　机器学习算法安全评估流程 11
7.1　流程要求 11
7.2　评估准备 11
7.3　评估方案 11
7.4　评估执行 12
7.5　评估结论 12
7.6　评估报告 12
附录 A（规范性）算法推荐服务安全要求 14
附录 B（规范性）算法推荐服务评估方法 21
参考文献 29

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。