ICS 35.030
CCS L 80

中华人民共和国国家标准

GB/T XXXXX—XXXX

信息安全技术
机器学习算法安全评估规范

Information security technology-Security specification and assessment methods for machine learning algorithms

（征求意见稿）

（本稿完成时间：2021 年 7 月 27 日）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX - XX - XX 发布　　XXXX - XX - XX 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。

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、清华大学、北京瑞莱智慧科技有限公司、国家信息技术安全研究中心、广州大学、国家计算机网络应急技术处理协调中心、华为技术有限公司、北京旷视科技有限公司、中国信息通信研究院、北京百度网讯科技有限公司、中国科学院信息工程研究所、阿里巴巴（北京）软件服务有限公司、深圳市腾讯计算机系统有限公司、北京奇虎科技有限公司、重庆邮电大学、深圳市大数据研究院、北京计算机技术及应用研究所、中国电子技术标准化研究院。

本文件主要起草人：上官晓丽、胡影、郝春亮、张宇光、苏航、胡嵩智、杨韬、景慧昀、张旭东、许晓耕、顾钊铨、吴月升、孟国柱、李实、付英波、梅敬青、王乐、董胤蓬、刘曦泽、王哲麟、赵芸伟、韩晗、张夏、彭骏涛、徐永太、张屹、徐雨晴、吴保元、韩磊、王秉政。

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在设计开发、验证测试、部署运行、维护升级、退役下线等阶段的安全要求和证实方法，以及机器学习算法的安全评估实施。

本文件适用于对机器学习系统中的算法进行安全评估，也适用于机器学习系统开发者和运营者在算法开发运营过程中进行自评估和改进安全措施。

2 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。

GB/T 25069—2010 信息安全技术 术语

GB/T 35273—2020 信息安全技术 个人信息安全规范

GB/T 37988—2019 信息安全技术 数据安全能力成熟度模型

3 术语和定义

GB/T 25069—2010 中界定的以及下列术语和定义适用于本文件。

3.1

对抗样本 adversarial examples

在数据集中通过故意添加细微的干扰所形成的输入样本，该类样本导致模型以高置信度给出错误地输出。

3.2

个人信息 personal information

以电子或者其他方式记录的能够单独或者与其他信息结合识别特定自然人身份或者反映特定自然人活动情况的各种信息。

[来源：GB/T 35273—2020，术语和定义 3.1]

注：个人信息的范围和类型参见 GB/T 35273—2020。

3.3

机器学习算法 machine learning algorithm

采用机器学习技术理论求解问题，明确界定的有限且有序的规则集合，并基于输入数据生成分类、推理、预测等的算法。

3.4

机器学习算法生命周期 machine learning algorithm lifecycle

机器学习系统的算法从起始到退役的整个演进过程，包括设计开发、验证测试、部署运行、维护升级、退役下线。

注：在机器学习算法生存周期中，某些活动可出现在不同的过程中，个别过程可重复出现。例如为了修复错误和更新，需要反复实施开发和部署过程。

3.5

算法失效 algorithm failure

算法丧失完成规定功能的事件。

3.6

野值 outliers

采样数据集合中严重偏离大部分数据所呈现趋势的数据。

3.7

准确率 accuracy

对于给定的数据集，正确分类的样本数占总样本数的比率。

4 缩略语

下列缩略语适用于本文件。

SDK：软件开发工具包（Software Development Kit）

5 概述

机器学习算法具有需要从数据或经验中学习、输出结果不确定、决策过程不可解释等特点，因此，机器学习算法安全不仅涉及算法自身安全，也涉及算法应用安全。机器学习算法生命周期各阶段均面临算法、数据和环境层面的安全风险。在所有场景中，机器学习算法安全需要满足保密性、完整性、可用性、可控性、鲁棒性和隐私性等基本安全属性。在公共服务、交通驾驶、金融服务、健康卫生、福利教育等重要领域，用于生命财产安全、个人权利保障等关键事项决策时如部署应用机器学习算法，对其安全要求则更为严格。

本文件中机器学习算法安全要求分为基本级与增强级：

——基本级：安全要求条款的字体未加粗，表示适用于所有场景的机器学习算法的安全要求；

——增强级：安全要求条款的字体加粗，表示在公共服务、交通驾驶、金融服务、健康卫生、福利教育等重要领域，用于生命财产安全、个人权利保障等关键事项决策时的额外安全要求。

本文件第七章所提证实方法与第六章所提安全要求是逐条对应的。与第六章中增强级安全要求对应的证实方法，在第七章中用字体加粗表示。

注1：机器学习算法安全属性及安全指标体系见附录A。

注2：机器学习算法安全风险见附录B。

注3：机器学习算法从数据或经验中学习的特性，使得算法安全性必须包括训练数据与测试数据的安全性。特别的，涉及到个人信息的机器学习算法还需要满足隐私性要求。

6 安全要求

6.1 通用

机器学习算法的通用安全要求包括：

a) 组织和个人开展机器学习算法相关的数据处理活动时，应具备基本数据安全能力，应达到 GB/T 37988—2019 中规定的 2 级水平，宜达到 3 级水平。

b) 相关组织和个人在开发或运营机器学习算法时，应确保所开发或运营的机器学习算法的保密性，确保机器学习算法模型、数据、依赖信息不被泄漏给未经授权的个人、实体或过程，包括但不限于：

1） 在验证测试阶段，通过算法的防范成员推理攻击和逆向攻击测试；

2） 数据/模型销毁应确保删除数据/模型无法恢复；

3） 进行访问权限设置，拒绝非授权主体访问。

注：保密性指标见附录 A.1。

c) 相关组织和个人在开发或运营机器学习算法时，应确保所开发或运营的机器学习算法的完整性，确保机器学习算法模型、数据、软硬件依赖信息等不被未经授权的方式替换或破坏，包括但不限于：

1） 在算法设计开发和部署应用阶段，对数据/模型/运行环境进行定期校验，防止数据/模型/运行环境被篡改；

2） 对数据集规模、均衡性、准确性、与算法任务相关程度等指标进行测试，确保其满足算法需求。

注：完整性指标见附录 A.2。

d) 相关组织和个人在开发或运营机器学习算法时，应确保所开发或运营的机器学习算法的可用性，确保一旦授权用户需要，就可以访问和使用机器学习算法模型、数据、依赖信息等，包括但不限于：

1） 修补公开的漏洞，阻止渗透等攻击手段干扰算法运行所依赖的软硬件正常运行；

2） 在设计文档中，对数据格式、大小等属性加以限制；

3） 测试算法失效事件发生后迅速恢复运行状态的能力，包括恢复时间与恢复程度；

4） 使用抽样检测与全部检测结合的手段，检验数据是否能准确表示其所描述的实际对象。

注：可用性指标见附录 A.3。

e) 相关组织和个人在开发或运营机器学习算法时，应确保所开发或运营的机器学习算法的可控性，确保在规定的条件下、在规定的时间内完成规定的功能，并且在机器学习算法行为失当时，提供机制使得算法运行行为可由操纵者接管的能力，包括但不限于：

1） 测试规定时间和条件下，算法失效的次数、发生故障的严重性和频繁程度、计算解决失效的比例，确保算法持续运行时间符合设定阈值；

2） 在部署运行阶段，记录算法正常服务时间、累计有效服务时间，并计算正常服务时间占比；

3） 在算法生命周期中，依据项目管理各阶段要求，编写文档材料与算法关键决策环节日志记录，提供可审计、可追溯能力。

注：可控性指标见附录 A.4。

f) 相关组织和个人在开发或运营机器学习算法时，应确保所开发或运营的机器学习算法的鲁棒性，确保机器学习算法系统在任何情况下均可以保持其性能水平，包括但不限于：

1） 利用压缩、损坏等有损数据和加入噪声、变换等干扰数据，测试算法功能实现正确性；

2） 分别在白盒、黑盒对抗场景下，测试算法准确率，并满足设定阈值；

3） 对算法进行对抗样本攻击、物理对抗攻击、后门攻击测试，确保攻击成功率低于设定阈值；

4） 在设计文档中设定模型反馈输出、查询次数、查询频率的阈值，避免访问控制攻击。

注 1：鲁棒性指标见附录 A.5。

注 2：对抗样本攻击见附录 C。

g) 相关组织和个人在开发或运营机器学习算法时，应确保所开发或运营的机器学习算法的隐私性，确保处理数据遵守法律和法规要求，保护个人信息和隐私，避免存储、泄漏敏感数据，包括但不限于：

1） 未经个人同意，不应使用其个人信息开展机器学习算法相关活动。法律法规规定无需同意的情况除外；

2） 对个人信息采用必要的数据脱敏措施。

注 1：个人信息包括人脸、声纹、身份证号等，见 GB/T 35273—2020。

注 2：隐私性指标见附录 A.6。

6.2 设计开发阶段

进行设计开发时，相关组织或个人：

a) 应对所使用的数据进行安全检测，对检测到的被污染数据进行修复或过滤，并保留检测处置记录；

注：检测包括数据变形检测、伪造检测、后门检测等。

b) 应根据使用场景的安全需求，分析确定算法可用性指标，保存分析记录，并确保算法开发符合指标设置；

c) 应根据应用场景的安全需求，分析确定以下训练数据指标，保存分析记录，并确保训练数据符合指标要求：

1） 训练数据规模阈值；

2） 训练数据均衡性指标；

3） 训练数据数据标注准确率阈值。

注：可通过多来源标注交叉验证等方式测算准确率。

d) 应设计应急处置机制，确保情况必要时算法能中断运行；

e) 应开展提高算法鲁棒性相关活动，包括但不限于：

1） 使用对抗训练等方法提高鲁棒性；

2） 记录鲁棒性改进过程，包括投入时间、重要操作等信息；

3） 评估算法鲁棒性提升效果并形成评估报告。

6.3 验证测试阶段

进行验证测试时，相关组织或个人：

a) 应采取动态测试与静态测试结合的方法，检测和定位算法缺陷、后门、潜在风险等；

b) 若委托第三方开展测试，应具有算法和数据保密的安全机制，可采用的方式包括但不限于：

1） 通过使用算法所有者提供的环境和设备开展测试；

2） 对同一算法使用两个或多个第三方对不同数据类型分别测试。

c) 应根据测试任务，分析设置测试数据与训练数据重复度指标，保存分析过程，并确保测试数据符合指标设置；

d) 应包含可复现性测试：

1） 应根据使用场景安全需求以及测试情况预先分析设置可复现性阈值，并保留分析过程；

2） 对算法可复现性进行验证测试，并记录测试结果。

注：可复现性是指算法处理相同数据产生相同或高度相似的结果。

e) 应根据测试任务，分析确定以下指标，保存分析记录，并确保测试数据符合指标要求：

1） 测试数据规模阈值；

2） 测试数据均衡性指标；

3） 测试数据数据标注准确率阈值；

4） 测试数据与测试任务相关性阈值；

f) 应根据测试任务，充分验证测试算法的鲁棒性，包括使用包含自然噪声、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进行测试；

g) 应适当模拟算法行为失当场景，检测算法是否提供人工接管、终止的方法或措施。

6.4 部署运行阶段

进行部署运行时，相关组织或个人：

a) 应对输入数据格式、大小等属性加以限制，防止特殊输入数据使模型出错的情况发生；在干扰性输入较多的场景时，应加入输入筛选过滤等机制确保算法稳定运行；

注：干扰性输入包括野值数据等。

b) 应通过限制模型的反馈输出、限制模型的查询次数、限制账号和 IP 的使用频率等方式保护模型内部参数等隐私数据，防止被攻击者逆向推测和还原模型参数、训练数据等；

c) 应采取措施降低算法的参数文件以及代码文件逆向风险，可采用手段包括加密存储算法、算法代码混淆等手段；

d) 应及时、准确、完整、清晰、无歧义地向使用者说明机器学习算法的作用、局限、安全风险和可能的影响，并解释相关应用过程及应用结果；

e) 应具有数据安全保护机制，保障数据的保密性、完整性、可用性，可采用的方式包括但不限于：

1） 加密算法；

2） 完整性校验。

f) 应设置应急处置机制，包括算法导致安全问题时可人工紧急干预、中止等；

g) 如部署应用不可解释的算法，应仅作为辅助决策手段，不作为直接决策依据。

6.5 维护升级阶段

进行维护升级时，相关组织或个人：

a) 若对算法进行大幅度调整、修改或升级，应对模型参数、配置进行及时更新，删除无关参数和数据，确保变更过程满足设计开发、验证测试、部署运行阶段的安全要求，并具有算法变更记录，至少包括算法变更的时间、描述，以及对应的模型参数、配置更新、删除的操作记录；

b) 应设置安全校验机制，对模型的升级包文件进行安全校验，确保升级包的安全性，并具有模型升级校验记录，至少包括模型升级校验的时间、版本以及关键校验信息的操作记录。

6.6 退役下线阶段

进行退役下线时，相关组织或个人：

a) 应设置退役下线的满足条件，并设置合理时间周期供销毁数据、模型；

b) 应对存储介质及相关文档材料中的数据，包括训练数据、测试数据、实例数据、特征数据、参数、算法输出等进行销毁；特征数据、参数和算法输出，同时满足以下条件时可不销毁：

1） 为进一步实现实际业务功能所必须；

2） 获得数据所有者明确授权同意；

3） 不涉及个人信息和有关部门认定的重要数据；

4） 经过混淆或加密处理，无法直接还原原始数据。

c) 应确保已销毁模型不能再被访问；

d) 若对个人信息去标识化，应进行删除或匿名化处理，确保涉及的个人信息不可被恢复；

e) 若对部署在多台设备上或通过云边协同的方式部署的模型实施退役下线，应从多台设备、云端等同时对模型及数据实施销毁。

7 证实方法

7.1 通用

机器学习算法安全通用要求的证实方法如下：

a) 确定其是否具有有效的认证材料，以证明该组织或个人具备 GB/T 37988—2019 中规定的等级 2 水平；

b) 机器学习算法保密性相关要求对应的证实方法如下：

1） 检查其在验证测试阶段，通过算法的防范成员推理攻击和逆向攻击测试；

2） 对销毁数据/模型进行恢复测试，确定其是否能被恢复；

3） 检查其是否对算法、模型、应用产品进行访问权限设置。

c) 机器学习算法完整性相关要求对应的证实方法如下：

1） 检查其是否在算法设计开发和部署应用阶段对数据/模型/运行环境进行定期校验；

2） **检查其是否对数据集规模、均衡性、准确性、与算法任务相关程度等指标进行测试，确定测试结果是否满足设计文档所提算法需求。**

d) 机器学习算法可用性相关要求对应的证实方法如下：

1） 检查其是否对公开漏洞进行修补；

2） 检查其是否在设计文档中对数据格式、大小等属性进行限制；

3） 检查其是否对算法失效事件发生后算法恢复运行状态所需的恢复时间与恢复程度进行测试；

4） 检查其是否对数据准确性进行测试，方法包括但不限于抽样检测与全部检测。

e) 机器学习算法可控性相关要求对应的证实方法如下：

1） 检查测试文档中是否记录了算法失效次数、发生故障的严重性和频繁程度、算法失效比例等测试数据，确认其测试结果是否满足阈值；

2） 检查其是否在算法部署运行阶段记录算法正常服务时间、累计有效服务时间、正常服务时间占比等指标；

3） 检查是否具有编写文档材料与算法关键决策环节日志记录。

f) 机器学习算法鲁棒性相关要求对应的证实方法如下：

1） 检查是否利用压缩、损坏等有损数据和加入噪声、变换等干扰数据对算法功能正确性进行测试；

2） 检查测试记录或第三方出具的测试报告，确定白盒、黑盒对抗场景测试准确率是否满足设定阈值；

3） **检查测试记录或第三方出具的测试报告，确定对抗样本攻击、物理对抗攻击、后门攻击测试成功率是否满足设定阈值；**

4） 检查设计文档，确定是否设定反馈输出、查询次数、查询频率的阈值。

g) 机器学习算法隐私性相关要求对应的证实方法如下：

1） 检查个人信息授权记录是否与数据规模一致，抽查个人信息及其对应的授权记录是否完备；

2） 检查在数据处理过程中是否对个人信息采取数据脱敏措施。

7.2 设计开发阶段

设计开发阶段，机器学习算法安全要求的证实方法如下：

a) 检查污染数据的检测记录，确定是否提供了对应的污染数据的检测和修复方法、修复结果，并测试 A.2 完整性指标训练数据一致性及测试数据一致性测试子项是否符合设计要求；

b) 访谈机器学习算法设计开发负责人，询问对算法的可用性需求，确定是否具有算法的可用性分析文档，测试 A.3 可用性是否符合设计要求；

c) 判定训练数据符合指标要求的证实方法：

1） 访谈数据集及机器学习算法设计开发负责人，检查训练数据集记录文档，测试 A.2 完整性指标数据准确性测试子项，查看数据集规模是否符合设计要求；

2） 访谈数据集及机器学习算法设计开发负责人，检查训练数据集记录文档，测试 A.2 完整性指标数据均衡性测试子项，检查设计文档中均衡性指标及均衡性指标合理性分析部分，检查数据集或数据统计材料确定是否符合设计要求；

3） 测试 A.2 完整性指标数据准确性测试子项，查看数据标注准确率是否符合设计要求，检查是否在算法训练过程中实施了数据校验和记录。

d) 访谈机器学习算法应急管理负责人，检查制度文件、设计文档，查看是否设置了事故应急处置机制、明确了事故处理流程、设置了事故信息回溯机制；

e) 访谈机器学习算法设计开发负责人，询问算法是否在公共服务、交通驾驶、金融服务、健康卫生、福利教育等重要领域，用于生命财产安全、个人权利保障等关键事项，检查鲁棒性评估记录，确定 A.5 抗攻击能力是否符合设计要求。

7.3 验证测试阶段

验证测试阶段，机器学习算法安全要求的证实方法如下：

a) 测试 A.5 鲁棒性指标数字世界白盒对抗鲁棒准确率、数字世界黑盒查询攻击对抗鲁棒准确率、数字世界迁移攻击对抗鲁棒准确率、物理世界对抗样本攻击成功率、模型后门攻击成功率等测试子项，检查上述指标是否符合设计要求；

b) 若委托第三方开展测试，检查测试文档中是否包含算法和数据保密的安全机制的记录；

c) 测试 A.2 完整性指标数据重复度测试子项，检查测试数据集与训练数据集的重复度是否符合设计要求；

d) 可复现性测试证实方法：

1） 检查测试文档，确定测试文档是否包括设置可复现性阈值，并保留分析过程；

2） 测试 A.3 可用性指标可访问性测试子项，测试 A.4 可控性指标有损数据鲁棒性、干扰数据鲁棒性等测试子项，检查评估指标体系中可访问性指标和可控性指标及对应指标测量值说明部分，检查使用测试数据集验证模型是否能对相同的数据能复现相同或相似度高的结果。

e) 判定测试数据符合指标要求的证实方法：

1） 测试 A.2 完整性指标数据规模测试子项，检查设计文档中数据规模指标及数据规模指标合理性分析部分，检查测试数据集或数据统计材料确定测试数据规模是否符合设计要求；

2） 测试 A.2 完整性指标数据均衡性测试子项，检查设计文档中均衡性指标及均衡性指标合理性分析部分，检查数据集或数据统计材料确定测试数据均衡性是否符合设计要求；

3） 测试 A.2 完整性指标数据准确性测试子项，查看数据标注准确率是否满足预先设定的阈值，检查算法测试过程中是否实施了数据校验和记录；

4） 测试 A.2 完整性指标数据任务相关性测试子项，检查设计文档中数据任务相关性指标及

数据任务相关性指标合理性分析部分，检查测试数据集或数据统计材料确定测试数据相关性是否符合设计要求。

f) 测试 A.5 鲁棒性指标数字世界白盒对抗鲁棒准确率、数字世界黑盒查询攻击对抗鲁棒准确率、数字世界迁移攻击对抗鲁棒准确率、物理世界对抗样本攻击成功率、模型后门攻击成功率等测试子项，检查上述指标是否符合设计要求；

g) 检查测试文档，确定是否包含检测算法人工接管、终止的方法或措施，检查模拟行为失当场景的合理性。

7.4 部署运行阶段

部署运行阶段，机器学习算法安全要求的证实方法如下：

a) 访谈机器学习算法负责人，检查设计及部署文档，测试 A.3 可用性指标数据质量测试子项，检查是否进行输入数据要求测试；测试 A.5 鲁棒性指标有损数据鲁棒性、干扰数据鲁棒性、数字世界白盒对抗鲁棒准确率、数字世界黑盒查询攻击对抗鲁棒准确率、数字世界迁移攻击对抗鲁棒准确率、物理世界对抗样本攻击成功率、模型后门攻击成功率等测试子项，检测上述指标是否符合设计要求；

b) 访谈部署运行负责人，检查设计文档和部署运行文档，测试 A.1 保密性系统信息保密性指标子项，测试 A.1 保密性数据保密性指标运行数据保密指标子项是否符合设计要求，测试 A.5 鲁棒性抗攻击能力指标测试子项，检查是否进行了访问控制测试；

c) 测试 A.1 保密性指标模型保密性测试子项，验证是否采取模型加密、代码混淆等手段防止模型的参数文件以及代码文件逆向；

d) 访谈部署运行负责人，检查相关设计运行文档，测试 A.3 可控性指标可审计性文档完备测试子项，检查是否进行了文档完备性测试；

e) 访谈机器学习算法负责人，检查设计运行文档是否记录数据安全保护机制部署情况，包括但不限于数据加密算法和完整性校验；

f) 访谈机器学习算法及应急管理负责人，测试 A.3 可用性可恢复性子项，检查设计文档和运行环境应急处置机制，检查算法的系统、产品或服务中设置事故应急处置机制部署情况，查看是否明确事故处理流程，确保在机器学习算法导致安全问题时具备人工紧急干预、终止等的能力；

g) 访谈机器学习算法负责人，查看设计及部署文档，测试 A.3 可控性指标可审计性文档完备测试子项，验证是模型决策或计算过程是否具有透明性、可解释性及可审计性，若具有不可解释性，确定是否将算法决策结果仅作为重要决策的辅助决策。

7.5 维护升级阶段

维护升级阶段，机器学习算法安全要求的证实方法如下：

a) 检查算法变更记录文件，测试 A.1 保密性指标模型保密性测试子项是否符合设计要求，测试 A.3 可用性指标算法准确性、软硬件依赖、计算资源可用性、恢复时间、可恢复能力、可访问性是否符合设计要求；

b) 检查模型升级校验记录文件，测试 A.2 完整性指标模型一致性测试子项，确定实际部署的模型与预先部署的模型是否相同，确定是否设置对模型升级包文件的安全校验机制并进行了校验。

7.6 退役下线阶段

退役下线阶段，机器学习算法安全要求的证实方法如下：

a) 检查算法相关设计运行文档，是否设置退役下线的满足条件，并设置合理时间周期供销毁数

据、模型；

b) 检查数据存储介质，是否将应删除数据进行销毁，检测留存数据是否经过混淆与加密。测试A.1 保密性指标数据保密性指标，使用数据恢复手段对数据存储介质进行检测，确认各设备中的已销毁数据是否能够被恢复；

c) 检查存储介质中的算法是否被销毁，测试 A.1 保密性指标模型保密性指标，使用数据恢复手段对算法存储运行介质进行检测，确认各设备中的已销毁算法是否能够恢复；

d) 测试 A.6 隐私性指标数据合规性、个人信息保护测试子项，检测是否进行了个人信息保护测试。

8 安全评估实施

8.1 安全评估形式

机器学习算法安全评估分为自评估和第三方评估两种形式，自评估和第三方评估互相结合、互为补充。

自评估是指机器学习算法开发者、运营者发起的对本组织机器学习算法等进行的安全评估，通常自评估由组织内部质量管理部门或测试部门发起，主要采用检查、测试形式自评估应在本标准的指导下，结合算法特定的安全要求实施。周期性进行的自评估可以在评估流程上适当简化。为保证安全评估的实施，涉及影响算法安全的相关方也应配合。

第三方评估可依据本标准的要求，实施完整的机器学习算法评估过程。第三方评估也可在自评估实施的基础上，对算法或相关内容实施评估。第三方评估由独立机器学习系统开发者和运营者等相关方的专业评估机构实施，主要采用访谈、检查、测试等相结合的形式。

8.2 安全评估准备

8.2.1 明确评估范围

选择评估实施所需的算法-模型文档及其他影响机器学习安全的说明文档，算法运行（原型）代码，数据，部署环境信息。

数据包括训练数据和生产数据，生产数据视评估运营主体的运营情况和数据主体的权限而定。

部署环境包括承载算法的网络边界，应用软件、计算及存储资源等。

8.2.2 评估工作启动

评估工作开展前应明确机器学习算法安全评估的背景、目标、原则和依据，充分调研待评估对象所属行业、领域相关标准及政策文件，组建评估团队，确定评估工作任务，第三方评估应与应用开发者、运营管理者签署保密协议。

8.2.3 评估方案制定

结合待评估对象的具体情况，编制评估方案，方案应包含：

a) 评估范围、对象、目标等；

b) 评估内容、实施方法、时间进度安排和使用的软硬件工具及环境，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；

c) 风险管控措施；

d) 人员安排、项目管理制度；

e) 被评估方需要配合的事项清单；

f) 被评估方应准备对应阶段评估所需的文档、代码及其他相关材料。

8.2.4 评估方案评审

对制定的评估方案的可行性、适用性及针对性进行评价。

8.3 安全评估执行

依据本文件第7部分所提证实方法进行逐条评价。

8.4 安全评估总结

8.4.1 综合分析

评估方对机器学习算法各生命周期的评估，应对评估过程中发现的安全问题和风险隐患进行分析评判，对机器学习算法的安全性予以综合分析评估。综合分析应主要包括：

a) 本标准保密性、完整性、可用性、可控性、鲁棒性、隐私性各个单元指标符合情况；
b) 本标准第 6 部分的不适用项及说明；
c) 部分符合或不符合指标项的风险问题，进行风险分析；
d) 评估发现的其他安全问题。
e) 针对部分符合或不符合指标项及其他安全问题提出的相关建议。

8.4.2 报告编制

评估方应就此次评估结果形成评估报告，评估报告包括但不限于评估对象描述、评估时间、算法是否符合评估要求的说明等。

8.4.3 结果反馈

评估方应将评估的工作情况和正式评估报告向被评估方反馈。

8.5 安全评估结果判定

机器学习算法安全评估的结果判定，包括以下三类：

a) 上述基本级要求中存在不满足项，判定为“不符合本标准基本级要求”；
b) 满足全部基本级要求，判定为“符合本标准基本级要求”；
c) 满足全部基本级和增强级要求，判定为“符合本标准增强级要求”。

附 录 A
（规范性）
机器学习算法安全评估指标体系

附录A给出了机器学习算法安全细粒度的评估指标体系，可针对不同应用场景的机器学习算法开展相关评估活动。

该指标体系由“属性—指标”两层组成，其中属性包括保密性、完整性、可用性、可控性、鲁棒性、隐私性等，由各属性对应的评估指标组成评估指标体系，以下给出了评估指标描述和细分项列表。

在实施评估过程中，应根据不同机器学习技术发展的成熟度和不同应用领域的安全需求选取相应的指标，并确保应用目标和使用方式符合国家法律法规、行业监管政策、标准规范以及伦理要求。

A.1 保密性指标

保密性指标用于评估机器学习算法生命周期各阶段的保密性，包括但不限于：

a) 数据保密性指标：评估未被授权者利用梯度信息、机器学习系统输出等信息，通过逆向工程等技术手段窃取机器学习相关数据的风险，包括训练数据保密、运行数据保密及抵御成员推理攻击的能力等，涉及数据传输、存储、计算、汇聚的保密性，个人信息保护及密钥安全等方面；

b) 模型保密性指标：评估模型的保密性，通过对被测系统持续访问，推测机器学习模型的参数或功能，判断所窃取或推断的训练数据与原数据的相似度；

c) 系统信息保密性指标：评估机器学习算法生命周期各阶段，软硬件依赖信息被窃取或非授权访问的情况。

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
1	数据保密性指标	训练数据保密	通过梯度信息、成员推理等方式，所窃取的训练数据与原数据的相似度	以查询次数为x轴，以仿制训练数据和被仿制训练数据的相似度偏差比例为y轴绘制的曲线，使用1-相似度计算	相似度越高，攻击成功率越高，则安全性越低	必选
2		成员推理攻击成功率	针对一组测试样本，推断其属于训练数据集的比例	针对一组测试样本，推断其属于训练数据集的比例	比例越低，则安全性越好	必选
3		运行数据保密	通过机器学习系统输出信息，所窃取的训练数据与原数据的相似度	以查询次数为x轴，以仿制训练数据和被仿制训练数据的相似度偏差比例为y轴绘制的曲线，使用1-相似度计算	相似度越高，攻击成功率越高，则安全性越低	必选
		数据销毁	应删除的数据是否未被销毁或仍能被恢复	检测是否能够直接提取或通过恢复手段，获取应删除的数据	如不能获取，则安全性较好；如能够获取，	必选

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
					则安全性较低	
4	模型保密性指标	模型窃取成功率	窃取的机器学习模型与原模型的相似度和性能	以查询次数为x轴，以仿制模型识别效果和被仿制模型识别效果的偏差比例为y轴绘制的曲线	曲线下面积越小，则安全性越好	必选
		模型销毁	应删除的算法是否未被销毁或仍能被恢复	检测是否能够直接提取或通过恢复手段，获取应删除的算法	如不能获取，则安全性较好；如能够获取，则安全性较低	必选
		模型防逆向	是否能够逆向出模型参数文件以及代码文件	检测是否应用了模型加密、代码混淆等防止模型被逆向的技术手段	技术手段运用越全面，则安全性越好	必选
5	系统信息保密性指标	系统信息保密	系统所采用的算法、数据、依赖等信息可以被非授权主体访问的程度	检测系统所采用的算法、数据、依赖等信息可以被非授权主体访问的次数	所能够获取的信息越少，则安全性越好	必选
6	其他指标	相关其他指标	相关其他指标满足保密性目标的程度	依据指标而定	依据测试指标而定	可选

A.2 完整性指标

用于评估机器学习生命周期各阶段的完整性，包括但不限于：

a) 数据完整性指标：评估算法所涉及数据的准确性和可靠性，从机器学习生命周期各阶段数据的完整、清晰、准确、可靠程度，包括数据一致性，数据均衡性，数据准确性等；

b) 运行环境完整性：评估算法依赖的运行环境的完整性水平，包括运行环境信息的一致性等；

c) 模型完整性指标：评估承载算法的 SDK 或产品被未经授权的方式替换或破坏的影响程度，衡量部署的模型与原始模型是否一致。

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
1	数据完整性指标	数据均衡性	包括数据类别均衡程度、无偏度	使用统计手段检测数据类别分布、偏度	数据类别均衡、无偏，则数据均衡性好	必选
2		数据准确性	数据所描述的实际对象真实值的程度	使用抽样检测与全部检测结合的手段，检验数据是否能准确表示其所描述的实际对象	检测样本中正确性比例越高，数据准确性越好	必选

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
3		训练数据一致性	衡量篡改后的训练数据与原始训练数据的一致程度	使用统计手段计算篡改后的训练数据与原始训练数据的比例	一致程度越高，则安全性越好	必选
4		测试数据一致性	衡量篡改后的测试数据与原始测试数据的一致程度	使用统计手段计算篡改后的测试数据与原始训练数据的比例	一致程度越高，则安全性越好	必选
5		数据重复度	衡量测试数据与训练数据重复程度	使用统计手段计算测试数据与训练数据重复的比例	数据重复度越低，则测试数据重复度越好	必选
6		数据任务相关性	衡量测试数据的任务相关程度	使用统计手段计算测试数据与训练数据分布的相关程度	数据任务相关性越高，则安全性越好	必选
7		训练数据规模	衡量训练数据的丰富程度	使用统计手段计算训练数据样本量	训练数据规模越大，则安全性越好	
8		测试数据规模	衡量测试数据的丰富程度	使用统计手段计算训练数据样本量	测试数据规模越大，则测试更加准确	
9		数据标注准确率	检测数据标注的准确程度	使用统计的方法检测正确标注的数据或全部数据之间的比例	标注准确率越高，则安全性越好	
10	模型完整性指标	模型一致性	衡量模型否一致	检测部署的模型与原始模型是否一致	如一致，则安全性较好	必选
11	运行环境完整性指标	运行环境信息一致性	衡量篡改后的运行依赖等系统信息与原始信息的一致性	通过代码单元测试、模块组件测试、集成测试等方法，测试算法所依赖各类运行环境的一致程度	一致程度越高，则安全性越好	必选
12	其他指标	相关其他指标	相关其他指标满足完整性目标	依据指标而定	依据测试指标而定	可选

A.3 可用性指标

用于评估机器学习生命周期各阶段的可用性，包括但不限于：

a) 可使用性指标：评估有效生存周期内算法可用程度、软硬件环境的依赖度及计算资源的可用的程度，包括算法准确性、可访问性、计算资源可用性、软硬件依赖等；

b) 可恢复性指标：评估算法和数据等从灾难状态恢复到可运行状态所需的时间和投入，包括恢复时间、可恢复能力等；

c) 训练数据质量指标：评估训练数据能准确表示其所描述的实际对象的程度。

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
1	可使用性指标	算法准确性	包括基于准确率，精确率，召回率、F-1值、PR曲线以及AUC等指标	针对算法特点，选取准确率，精确率，召回率、F-1值、PR曲线以及AUC等指标，评估算法可用性。	各项指标值越高，算法可用性越强	必选
2		软硬件依赖	机器学习生命周期中所需软硬件被攻击对可用性的影响程度	基于公开的漏洞，使用渗透等攻击手段，干扰算法运行所依赖的软硬件的正常运行	软硬件依赖度越低，安全性越高	必选
3		计算资源可用性	评估CPU、GPU等计算资源被攻击对于可用性的影响程度	使用针对CPU、GPU等计算资源的攻击手段，计算不同攻击下，算法运行的情况	抗干扰能力越强，安全性越高	必选
4		可访问性	检测一定时间内，算法需要被访问时，其运行结果的可获取程度	使用抽样方式，计算一段时间内，算法可被访问的有效次数和比例	指定有效次数下，可访问的比例越高，则安全性越好	必选
5		输入数据要求	包括是否对数据格式、大小等属性加以限制	利用不同格式、大小等属性的数据进行测试，判断是否存在限制逻辑	对输入数据属性进行限制，则安全性越好	必选
6	可恢复性指标	恢复时间	在安全事件发生后，从算法失能失效导致业务停顿之刻开始，到算法恢复至可以支持智能系统运行之刻为止，此两点之间的平均时间	规定时间和条件下，模拟安全事件，统计算法的恢复时间平均值、最大值、众数和中位数	恢复时间各项统计指标越低，安全性越高	必选
7		可恢复能力	在安全事件发生后，由算法失效至修复可用后的恢复有效程度	规定时间和条件下，模拟安全事件，统计算法的恢复到完备功能的比例	比例越高，安全性越高	必选
8	数据质量指标	数据质量	包括准确性（数据准确表示其所描述的实际对象真实值的程度）、规范性（数据符合数据标准、数据模型、业务规则、元数据或权威参考数据的程度）等	使用抽样检测与全部检测结合的手段，检验数据是否能准确表示其所描述的实际对象	检测样本中正确性比例越高，数据准确性越好	必选
9	其他指	相关	相关其他指标满足可	依据指标而定	依据测试指	可

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
	标	其 他 指标	用性目标		标而定	选

A.4 可控性指标

用于评估机器学习生命周期各阶段的可控性，包括但不限于：

a) 耐久性指标：评估规定时间与条件下，持续运行时发生故障的严重性和频繁程度，包括算法持续运行情况等；（增加：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等；应明确事故处理流程，确保在人工智能安全风险发生时作出及时响应，如停止问题产品生产、召回问题产品等；设置事故信息回溯机制）；

b) 可审计性指标：评估管理过程文档完备程度及关键环节的可追溯、可审计能力，包括文档完备程度等；

c) 可持续运行性指标：评估算法正常运行能力，包括正常服务时间等；

d) 容错性指标：评估算法的避免算法失效及容错的能力，包括失效密度、失效解决率等。

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
1	耐久性指标	持 续 运行	在规定的条件下，在规定时间内持续运行时发生故障的严重性、频繁程度和容错性	规定时间和条件下，统计算法发生故障的严重性和频繁程度，依照容错测试样例，统计容错策略边界	持续运行时间越多，安全性越好	必选
2	可审计性指标	文 档 完备	具有完备的文档材料与算法关键决策环节日志记录	依据项目管理各阶段要求，文档资料齐备，评估其可审计、可追溯能力	资 料 越 完备，安全性越高	必选
3		累 计 有 效 服 务 时间	提供的有效服务时间总和	计算指定测试时间段有效服务时长	指 定 时 间 内，服务时长越长，安全性越高	必选
4		正 常 服 务 时间	在规定的条件下，在规定时间内算法的正常使用时间	规定时间和条件下，统计算法正常服务的时间	正常服务时间越长，安全性越高	必选
5		正 常 服 务 时 间 占比	评估机器学习系统正常提供服务时间段占总服务时间段的比例	规定时间和条件下，基于测试数据集的输入数据，计算正常提供服务时间段占总服务时间段的比例	服务时间占比越高，则安全性越好	必选
6	容错性指标	失 效 密度	在一定的试验周期内，计算算法实效的情况提，提示信息不正确、模型未按预期要求进行动作等都属于失效	计算一定的试验周期内的算法的失效次数	失 效 数 越 小，安全性越好	必选

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
7		失效解决率	检测到的失效中已经解决的失效比率	计算一定的试验周期内算法实效后，定位安全问题并解决的比例	在失效数不为零时，比率越高安全性越好；失效数为零时，不计算该项结果	必选
8	其他指标	相关其他指标	相关其他指标满足可控性目标	依据指标而定	依据测试指标而定	可选

A.5 鲁棒性指标

用于评估机器学习生命周期各阶段的鲁棒性，包括但不限于：

a) 正确性指标：评估算法面对非正常数据输入时的正常运行能力，包括抵御干扰数据、有损数据等的情况；

b) 抗攻击能力指标：评估算法面对攻击时正常运行的能力，包括抵御数字世界攻击、物理世界攻击、黑白盒攻击、后门攻击等的能力。

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
1	正确性指标	有损数据鲁棒性	根据测试结果计算算法功能实现正确性	利用压缩、损坏等有损数据进行测试，根据测试结果计算算法功能实现正确性	算法功能实现正确性越高，则鲁棒性越好	必选
2		干扰数据鲁棒性	根据测试结果计算算法功能实现正确性	利用加入噪声、变换等干扰数据进行测试，根据测试结果计算算法功能实现正确性	算法功能实现正确性越高，则鲁棒性越好	必选
3	抗攻击能力指标	数字世界白盒对抗鲁棒准确率	计算白盒对抗场景下，算法识别的准确性	以生成对抗样本的扰动大小、攻击迭代轮数为x轴，以对应模型的识别准确率为y轴而绘制的多条曲线	选取x轴中某些点，计算曲线在这些点的取值并取平均，平均值越高，则安全性越好	必选
4		数字世界黑盒查询攻击对抗	计算黑盒对抗场景下，算法识别的准确性	以生成对抗样本的扰动大小、查询次数为x轴，以对应模型的识别准确率为y轴而绘制的多条曲线	选取x轴中某些点，计算曲线在这些点的取值并取平均，平均值越	必选

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
		鲁棒准确率			高，则安全性越好	
5		数字世界迁移攻击对抗鲁棒准确率	计算迁移性攻击场景下，算法识别的准确性	以生成对抗样本的扰动大小为x轴，以对应模型的识别准确率为y轴而绘制的多条曲线	选取x轴中某些点，计算曲线在这些点的取值并取平均，平均值越高，则安全性越好	必选
6		物理世界对抗样本攻击成功率	计算物理对抗场景下，攻击成功的比例	生成物理世界对抗样本，根据攻击结果计算攻击成功率	攻击成功率越低，则安全性越好	必选
7		模型后门攻击成功率	计算模型后门攻击下，攻击成功的比例	尝试植入后门并在输入样本上叠加触发器，根据攻击结果计算攻击成功率	攻击成功率越低，则安全性越好	必选
8		访问控制	是否采取限制手段来控制攻击者能够获取的信息	检测是否采用了限制模型的反馈输出、限制模型的查询次数、限制账号和IP的使用频率等访问控制手段		
9	其他指标	相关其他指标	相关其他指标满足鲁棒性目标	依据指标而定	依据测试指标而定	可选

A.6 隐私性指标

用于评估机器学习生命周期各阶段的隐私性，包括但不限于：

a) 合规性指标：评估数据生命周期各阶段的合规程度，包括数据合规性等；

b) 防护性指标：评估机器学习生命周期各阶段数据隐私的防泄漏、安全防护水平，包括数据防泄漏、银行学、个人信息保护等。

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
1	合规性指标	数据合规性	数据采集等环节合法合规的程度	依照现有个人信息保护、数据安全政策法规、	合法合规程度越高，安	必选

序号	指标类别	测试子项	测试子项检测目标	测试方法	测量值说明	类型
				标准规范进行评估	全性越好	
2	防护性指标	防泄漏	基于梯度信息，所还原隐私数据与原数据的相似度	使用梯度信息攻击方法，统计所还原隐私数据与原数据的相似度和比例	相似度及比例越低，则安全性越好	必选
3		数据隐含性	评估机器学习模型记住或揭示训练数据的可能性	使用逆向手段，统计机器学习模型推断的数据比例	比例越小，则安全性越好	必选
4		个人信息保护	评估算法各阶段是否对个人信息采取保护措施	采用技术、文档及现场查阅等方式，查看是否对关键场景采取了同态加密、传输通道加密等个人信息保护手段和措施	脱敏数据越多，脏数据越少，则安全性越好	必选
5	其他指标	相关其他指标	相关其他指标满足可用性目标	依据指标而定	依据测试指标而定	可选

A.7 指标测算方式

机器学习算法安全评估指标的测算方式，包括以下三类：

a) 专家评估：对于可审计、合理性等指标的测算，可以通过组织行业内专家进行综合评估；
b) 数据集测算：对于可用性、鲁棒性中意外故障等指标，使用测试数据集进行测算；
c) 模拟攻击测算：对于保密性、鲁棒性中有意动机等指标，使用模拟安全攻击方式进行现场测算。

A.8 指标测算和发布要求

机器学习算法安全评估指标的计算依赖于安全攻击方法、安全测试数据集的选择和构建。为保障安全评估指标的透明性和公平性，对于机器学习算法安全评估指标的计算、发布应遵循以下要求：

a) 对通过衡量机器学习算法在安全攻击下的性能表现来计算的安全评估指标，应优先选择当前攻击性能较好的多种安全攻击方法进行综合评价，并在计算和发布时应明确并公开所使用的安全攻击方法以及关键参数设置；
b) 对通过衡量机器学习算法在测试数据集上的性能表现来计算的安全评估指标，应在计算和发布时，明确并公开所使用的测试数据集规模、测试数据种类、典型测试数据样例等关键信息。

附 录 B
（资料性）
机器学习算法安全风险

B.1 机器学习算法分类

机器学习算法是计算机基于数据做出和改进预测或行为的一套方法。依据训练样本包含的信息以及反馈方式、训练方式、构建方式的不同，可以对机器学习算法进行不同分类。

依据训练样本包含的信息以及反馈方式的不同，机器学习算法分为监督学习、无监督学习和强化学习三类。其中，监督学习依据处理任务的不同，可以分为回归算法和分类算法两类；无监督学习依据处理任务的不同，可以分为聚类、降维和关联分析；强化学习依据学习方式的不同，可以分为有模型学习和免模型学习。

依据训练方式的不同，机器学习算法可以分为批量学习和在线学习两类。

依据算法构建方式的不同，机器学习算法一般可以分为使用第三方现有算法进行重新训练得到算法，以及原创设计训练得到的算法两类。

依据学习原理的不同，机器学习算法一般可以分为符号主义学习、连接主义学习、统计学习三类。其中符号主义学习算法的典型代表包括决策树、基于规则的学习等。连结主义学习算法主要是指感知机、BP网络等浅层神经网络以及深度神经网络。统计学习算法主要包括支持向量机、贝叶斯分类、主成分分析等。

依据算法训练方式的不同，机器学习算法一般可以分为在线训练和离线训练两类。

依据算法不当应用对于数字世界、物理世界和人类社会产生安全危害的不同，机器学习算法一般可以分为仅影响数字世界安全的机器学习算法，影响物理世界安全的机器学习算法，影响人类社会安全的机器学习算法三类。其中仅影响数字世界安全的机器学习算法，包括在信息通信、休闲娱乐等领域应用的机器学习算法。影响物理世界安全的机器学习算法，包括在石油、化工、制造、农业等领域应用的机器学习算法。影响人类社会安全的机器学习算法，包括在军事、金融、医疗、交通等领域应用的机器学习算法。

B.2 机器学习算法脆弱性与攻击威胁

B.2.1 机器学习算法脆弱性

机器学习技术局限、机器学习算法设计有误、机器学习算法软件缺陷、数据安全管理缺失或管理不严、机器学习框架安全漏洞等因素，均会引发机器学习算法安全脆弱性。包括：

a) 机器学习技术局限：机器学习算法存在弱鲁棒性、不可解释性、偏见歧视等尚未客服的技术局限。弱鲁棒性是指机器学习算法在面对复杂多变的实际应用场景、非正常恶意干扰等情况时可能产生意外非正确结果。不可解释性是指人类无法理解基于深度神经网络的机器学习算法的内部运行逻辑、训练得到的参数含义、决策产生原因等方面内容，为机器学习算法的问题定位、调试修改、责任追查等带来挑战。偏见歧视是指机器学习算法从带有社会已存在歧视的训练数据集中自主学习问题解决方案，将产生潜藏偏见的决策结果；

b) 机器学习算法设计有误：机器学习算法目标函数、求解方式等设计有误，无法实现设计者预设目标，导致产生偏离预期的不可预测、不可控行为；

c) 机器学习算法软件缺陷：机器学习算法在设计研发、维护升级等阶段未实施有效软件质量管理，带来安全漏洞；

d) 数据安全管理缺失或管理不严：在机器学习算法全生命周期中，未严格按照国家法律法规和标准规范要求，对机器学习算法的训练及数据的收集、保存、使用、委托处理/共享/转让/公开披露等环节进行管理，带来用户隐私泄露、数据预处理不规范等多方面风险；
e) 机器学习框架安全漏洞：机器学习框架及其使用依赖的第三方库自身存在安全漏洞和后门，将为基于此开发的机器学习算法带来相应的安全隐患。

B.2.2 机器学习算法攻击威胁

针对机器学习算法的新型安全攻击威胁包括对抗样本攻击、数据投毒攻击、算法后门攻击、模型窃取攻击、隐私窃取攻击等：

a) 对抗样本攻击：机器学习模型在运行阶段会受到对抗样本的攻击，对抗样本是指攻击者向正常样本中恶意添加微小的、人眼不可见的噪声，导致模型发生错误的样本。由于机器学习模型可能会部署在不受控制的真实场景中，对抗样本攻击成为了一种实际的安全风险；
b) 数据投毒攻击：机器学习模型在训练阶段会受到数据投毒的攻击，数据投毒是指攻击者向训练数据中添加一部分恶意构造的样本，使得训练得到的模型存在安全问题，比如模型的预测出现偏斜攻击、减低模型准确率、插入后门等等。数据投毒攻击的形式包括直接修改训练数据、利用反馈误导等。反馈误导攻击；
c) 算法后门攻击：算法后门攻击是指向机器学习模型中插入后门，被后门攻击的模型在正常的数据中表现良好，但是遇到特定的数据时会出现不合理的错误预测，以达到攻击者的目的。向模型中插入后门的方式包括修改训练数据和修改模型参数；
d) 模型窃取攻击：机器学习模型在运行阶段会受到模型窃取攻击，攻击者希望通过使用对模型的访问获得的信息构建替代模型，以盗取被攻击模型的特定功能；
e) 隐私窃取攻击：机器学习模型所利用的训练数据也面临隐私被窃取地风险，攻击者希望通过成员推理（判断某个数据是否在训练数据集中）、数据逆向还原（还原训练数据）、属性推理（判断某个特征是否用于训练模型）等方式获得模型使用训练数据的某些信息。数据隐私窃取攻击会导致用户的敏感信息被恶意使用。

B.3 设计开发阶段的安全风险

设计开发阶段主要用于明确任务、采集并形成数据集、设计相应的机器学习算法、选择机器学习框架等，该阶段，面临模型、数据、环境层面的安全风险。

B.3.1 算法层面的安全风险

在设计机器学习算法和模型时，需要考虑正确性等问题，模型层面面临的安全风险包括：

a) 模型的正确性无法满足任务需求地风险：算法的正确性是指算法能够生成行为符合预期的机器学习模型。由于不同的任务对于正确性的需求不同，当设计的机器学习算法无法保证符合任务需求的正确性时，该模型在实际应用时将引发安全问题；
b) 模型运行效率无法满足任务需求地风险：模型运行效率是指算法执行时花费的时间。当设计的算法需要花费较长时间，无法满足任务的效率需求时，如设计的算法时间复杂度是指数级，无法满足任务场景的秒级相应等需求，此类算法自身将成为一种安全风险；
c) 算法鲁棒性风险：由于机器学习模型的训练无法遍历所有目标领域的数据，使得训练得到的模型对外界环境干扰以及恶意攻击的鲁棒性较差。攻击者能够利用模型的这一脆弱点较容易的构造对抗样本使模型结果出错，从而导致严重安全隐患。

B.3.2 数据层面的安全风险

机器学习主要通过数据集进行训练，从数据集中学习相应的模式，并通过多种指标对学习效果进行评价。而数据集规模、数据集均衡性、数据标注的准确性、数据表示形式的规范化、数据是否被污染、隐私数据泄露等将成为设计开发阶段的安全风险。包括：

a) 数据集规模不足地风险：不同的机器学习任务需要的数据集规模不同，当数据集规模不够大，不足以支撑机器学习算法的有效学习时，会导致机器学习算法达不到和具体任务相对应的准确度要求，从而影响模型执行时的安全性；
b) 数据集均衡性差地风险：数据集均衡性用来描述数据集中包含不同类别的样本数量。当数据集均衡性差，某些类别样本数量少时，机器学习算法将难以有效学习多种类别的特点，将严重影响机器学习针对部分类别识别的准确性。此外，不均衡数据用于算法学习，将引发模型公平性风险；
c) 数据标注不准确地风险：对于有监督学习类别的算法，数据标注是数据集的重要组成部分，也是影响机器学习安全性的重要因素。然而，大多数数据集的标注是由人工众包形式完成，很难保证所有参与标注的人员都能正确标注，如果出现错标、漏标等情况，将会导致算法很难正确学习，引发错误分类或预测的问题；
d) 数据表示形式不规范地风险：在智能驾驶、智慧金融、智慧医疗等应用场景中，原始数据可能包含大量与任务无关的信息，数据表示形式不恰当或者不规范，将造成机器学习算法无法准确学习关键特征，影响算法的准确性，可能造成算法效率大幅度下降；
e) 数据集被污染地风险：设计开发阶段采集的数据集可能面临数据被恶意污染地风险，即数据投毒。由于恶意行为修改数据集中的数据，造成机器学习算法学习到错误的特征，引发模型出错；
f) 数据过度采集导致隐私数据泄露地风险：个人信息未加密或未脱敏使用时，攻击者很容易从模型中提取数据，造成数据集中个人信息的泄露。

B.3.3 环境层面的安全风险

机器学习模型的环境依赖其选择的框架及第三方库，而框架及第三方库自身存在漏洞会导致模型运行错误等风险：目前机器学习算法的设计和实现往往依赖开发框架的支持，这些框架本身往往使用大量的第三方库，在设计开发阶段使用的框架、第三方库等本身存在的安全漏洞和后门会导致基于开发的机器学习算法也存在相应的安全隐患。此外，操作系统、硬件架构和硬件配置可能引入风险，比如兼容性问题、处理精度问题、安全性问题、计算能力等。

B.4 验证测试阶段的安全风险

验证测试阶段是对涉及开发阶段获取模型功能和安全性的验证和测试，当算法经过数据训练得到具体的模型时，该阶段，旨在验证模型的功能、性能等是否满足事先定义的需求，在实际开发过程中，如果不满足则回退到设计开发阶段。该阶段，面临模型、数据、环境层面的安全风险。

B.4.1 算法层面的安全风险

a) 模型正确性低地风险：不同任务的正确性需求不同。当模型在测试数据集上的正确性无法达到需求，需要重新进行开发；
b) 模型运行效率低地风险：不同任务的效率要求不同。当模型在测试数据集上的效率无法达到需求，需要重新进行开发，优化算法；
c) 模型泄露隐私数据地风险：算法存在可能暴露数据集、通过简单测试办法可获取隐私数据等问题；
d) 算法公平性不足地风险：算法做出存在偏见和违反公平性原则的输出；
e) 算法和模型鲁棒性差地风险：当测试数据中包含较小的偏差或部分非正常分布的数据，算法的

结果受到较大影响，输出发生很大的变化；

f) 模型泛化能力差地风险：当使用和实际场景、任务相关的数据集测试时，模型的泛化能力差，将很难输出正确的结果。

B.4.2 数据层面的安全风险

在验证测试阶段，将通过大量的测试数据验证模型的准确性、效率、鲁棒性等，该阶段，面临测试数据和训练数据高度重复、测试数据集不足、均衡性差等安全风险。包括：

a) 测试数据和训练数据重复度高地风险：当用来测试的数据和训练数据集重复度高时，模型将在测试数据上表现优异，但是无法验证其在未知测试数据集上的性能；

b) 测试数据集规模不足地风险：测试数据如果规模较少，将无法对可能遇到的各种特殊数据进行有效验证，导致模型在特殊数据上判断出错；

c) 测试数据集均衡性差地风险：测试数据均衡性不足，部分类别的测试数据少时，将很难对模型在各类数据上的表现进行验证；

d) 测试数据集合实际任务相关性不足地风险：当未采集和实际应用场景、任务相关的数据进行测试时，将导致模型在实际运行过程中无法正确处理实际数据的问题。

B.4.3 环境层面的安全风险

机器学习算法依托特定的框架进行实现，在验证测试阶段，如果未对框架及第三方库自身的漏洞进行有效测试，将导致模型在运行过程中出现错误。

B.5 部署运行阶段的安全风险

机器学习部署运行阶段是在形成机器学习模型之后，将模型部署于实际应用的过程。该阶段，可能面临实际数据未知、框架更新、模型被攻击、运行环境适配等问题。该阶段，主要面临模型、数据、环境层面的安全风险。

B.5.1 算法层面的安全风险

a) 模型运算效率低地风险：在验证测试阶段，机器学习模型的效率已被检测，但在实际部署运行中，真实环境对于效率的要求可能会随着时间发生改变，没有及时更新的模型可能无法满足真实环境的效率要求；

b) 模型后门风险：攻击者在机器学习模型中植入特定的后门，使得模型虽然对正常输入与原模型判断一致，但对特殊输入的判断会受攻击者控制，从而造成模型的输出出现特定的错误；

c) 模型隐私风险：攻击者可以通过公共访问接口对模型进行多次访问，根据输入与输出的映射关系，在没有得知模型参数的情况下，构建出与被攻击模型相似度很高的模型，逆向推测并还原模型的参数信息等；

d) 数据反演导致数据泄露地风险：利用模型接口调用返回的信息进行反演攻击还原训练数据或部分隐私数据的安全风险；

e) 注入攻击风险：攻击者利用机器学习模型设计上的安全漏洞，将恶意的命令注入模型中，当实际数据满足预先设定的触发条件时，会造成模型完成注入的攻击行为；

f) 拒绝服务攻击风险：攻击者利用模型缺陷或者程序漏洞，通过构造特殊数据、利用敏感数据攻击模型，造成机器学习服务崩溃或者系统内存溢出等拒绝服务；

g) 可解释性风险：机器学习模型常被应用于医疗、收入预测、个人信息评估等安全敏感领域，如果模型输出结果的逻辑缺乏可解释性，易造成人们对于模型有效性的质疑，甚至反对；

h) 算法鲁棒性风险：新的算法仍然无法覆盖数据的可变空间，无法对恶意输入的样本做出正确的

判断。攻击者能够利用算法的这一特点构造对抗样本、恶意样本来攻击算法，导致模型出错。

B.5.2 数据层面的安全风险

当模型实际部署后，可能遇到实际环境中野值数据、带自然噪声扰动的数据、特殊扰动数据、数据被污染等风险，包括：

a) 干扰数据风险：实际环境中存在一些极端的输入数据与其余输入数据的差异较大，此类野值数据可能造成模型的输出出现错误；
b) 自然噪声扰动风险：在实际环境中，正常的输入数据会受到环境因素的影响，输入可能携带着无法预知的自然噪声扰动，将导致已部署的机器学习模型出现偏差；
c) 特殊扰动攻击风险：通过某种特殊方式修改输入，使得机器学习模型输出错误信息。此类特殊扰动攻击包括对抗样本攻击等；
d) 训练数据被污染风险：部分在线学习或演进学习的机器学习模型，部署在环境以后会根据实际的数据进行在线训练，自适应地调整模型参数等，攻击者在该过程中通过数据投毒，使用污染的数据训练此类模型，将造成模型的输出逐渐出现错误；
e) 数据集分布迁移风险：模型通常假设训练数据和真实数据服从相同分布，但模型部署在实际应用中时，数据集分布可能会发生迁移，即真实数据集分布与训练数据集分布之间存在差异性，从而造成模型输出产生偏差；
f) 数据泄露风险：机器学习模型部署以后，攻击者可能通过反复调用、查询模型，根据模型返回的信息还原训练数据，从而造成数据泄露等风险。

B.5.3 环境层面的安全风险

a) 机器学习框架更新风险：大部分机器学习模型是基于特定的框架开发的，当框架、第三方库更新时，若不及时对部署在此框架上的模型进行相应的调整，模型将会存在配置问题等风险；
b) 软硬件平台的部署风险：真实部署的环境中，操作系统、硬件架构和硬件配置等可能造成风险，如兼容性问题，处理精度问题，安全性问题，计算能力问题；
c) 供应链风险：攻击者在机器学习模型供应链过程中，通过逆向破解机器学习模型、通过控制软件/硬件渠道注入恶意代码等，造成恶意代码的传播。

B.6 维护升级阶段的安全风险

在机器学习算法或模型部署后，由于业务需要，会对算法和模型进行更新升级。在这个过程中，当算法和模型更新时，需要考虑在算法开发、验证测试、部署运行阶段相应地风险；当模型升级需要根据数据重新训练时，可能面临数据被污染等风险；当框架更新时，算法和模型需要及时升级，当软硬件等运行环境改变时，也需要对应进行维护。该阶段，主要面临模型、数据、环境层面的安全风险。

B.6.1 算法层面的安全风险

a) 当算法和模型进行大幅度调整、修改、升级时，需要重新考虑新的算法和模型在设计开发、验证测试、部署运行阶段面临的安全风险；
b) 模型参数未及时更新地风险：算法和模型升级时，可能遇到模型参数更新不及时、模型参数未正确删除等风险；
c) 模型配置冲突地风险：模型升级时，新旧模型的配置不一致，配置未正常升级将导致新模型无法正常运行等风险。

B.6.2 数据层面的安全风险

在维护升级阶段，部分机器学习算法和模型可能需要根据数据进行再次训练，在这个过程中将面临数据被污染地风险，即攻击者可能通过数据投毒，影响算法和模型的正确性。

a) 数据质量风险：在机器学习算法维护升级阶段，往往补充数据对当前算法版本进行优化。新收集数据集的规模、数据表示方式、数据均衡性和数据标注质量，都可能影响机器学习模型的内生安全；
b) 数据投毒攻击：在机器学习算法维护升级阶段，特别是在算法的再训练过程中，可能会引入数据投毒风险，人为对部分训练集数据进行篡改，会直接误导训练过程，导致模型训练过程不成功，或者导致模型被注入后门，无法对误导样本产生正确的预测结果，使得模型算法出现问题；
c) 数据隐私风险：新数据的加入，或者数据的替换和更新，都涉及数据的存储过程，在此过程中需要保证数据的可控性，规避数据被人为操作、破坏。

B.6.3 环境层面的安全风险

a) 和部署运行阶段类似，当模型由于框架更新、框架变更、第三方库更新时，若不及时对部署在此框架上的模型进行相应的调整，模型将会存在配置问题等风险，可能导致模型无法正常使用；
b) 在维护升级过程中，操作系统、硬件架构、硬件配置更新，但是模型未及时更新，将导致兼容性问题，例如模型处理的精度不足、计算能力不够等问题。

B.7 退役下线阶段的安全风险

退役下线阶段需要对相应的数据、模型等进行销毁，该阶段，主要包括模型和数据层面的安全风险。

B.7.1 算法层面的安全风险

a) 模型和算法泄露地风险：由于模型销毁过程的不当，可能造成模型、算法泄露地风险；
b) 模型参数泄露地风险：由于模型参数存储方式的不同，不同模型的销毁方法不同，可能造成模型参数保留在内存中，导致参数泄露地风险；
c) 模型未完全销毁地风险：模型涉及的配置文件、模型参数等众多，可能遇到未完全销毁模型地风险；
d) 多台设备未同时销毁模型地风险：部分模型同时部署在多台设备上，以及通过云边协同的方式部署，销毁时可能遇到未完全销毁地风险；
e) 误删或恶意删除模型地风险：由于权限设置不当，可能遇到误删或人为恶意删除模型地风险。

B.7.2 数据层面的安全风险

a) 数据未彻底销毁的风险：训练数据、测试数据、实际场景中的数据可能保留在内存、备份硬盘中，导致数据未彻底销毁；
b) 隐私数据泄露的风险：数据销毁过程中，可能由于销毁手段和强度不够，导致隐私数据泄露地风险；
c) 数据被误删或恶意删除的风险：由于数据权限设置不当，部分实际采集的数据、训练数据可能由于误删或者人为恶意删除，造成模型难以复现的问题。

附 录 C
（资料性）
对抗样本攻击

C.1 对抗样本

对抗样本指的是一类人为构造的样本，通过对原始样本添加特定的扰动，使得分类模型对新构造的样本产生错误的分类判断。现有的许多机器学习算法都很容易受到对抗样本的攻击。对抗样本产生的本质原因是目前的机器学习模型大多是做机械的数据拟合，无法像人类一样理解所要执行任务的要素，在训练样本无法覆盖所有样本空间的情况下，造成训练的模型边界与真实决策边界不一致，形成对抗样本的空间。

C.2 对抗攻击的目标

机器学习模型对抗攻击的重要目标就是在微小的扰动下，造成模型输出误差。根据最终的难度和影响大小，可以大致分为以下四种情况：

a) 无目标误预测：使得模型输出错误的预测结果，也就是通过增加扰动使得模型的输出和预设的输出不同；
b) 有目标误预测：通过增加扰动使得模型的输出为指定的错误结果；
c) 有源目标误预测：通过增加扰动，使得模型对于指定输入实现有目标误预测。

C.3 对抗攻击的类型

对抗攻击是指在模型使用过程中，攻击者通过对模型增加扰动，实现特点攻击。根据攻击者能够获得的模型信息，可以分为以下几种不同类型：

a) 白盒攻击:能够获取到机器学习模型的网络结构、网络的权重参数、用于训练模型的训练数据。也就是说攻击者能获取到这个模型的几乎所有数据,包括损失函数、最终模型训练得到的参数、训练的方式等。在白盒攻击的场景下，使用较小的扰动，模型被欺骗的概率仍可以达到 100%，而且大部分的防御方法仍然不能解决机器学习模型/算法在白盒攻击场景下的安全性问题；
b) 灰盒攻击：攻击者能够获取到模型的基本结构，但是不知道模型的具体参数；或者可以获取用于训练的样本数据，也就是能够知道模型训练集的分布，但是并不知道模型具体的结构；
c) 黑盒有查询攻击：在通常的应用中，机器学习模型/算法是不会暴露给攻击者的，但是攻击者通常可以得到模型对于任意输入的预测结果。例如对于各个公司所开发的人脸识别的 API，用户可以将图片上传到模型所在的服务器上，这些 API 会返回预测的结果。在这种场景下，攻击者并不能够获取目标模型，但是可以通过访问的方式估计出目标模型的运行方式，以访问的结果为经验逐渐增强攻击的效果，从而实现黑盒攻击；
d) 黑盒无查询攻击：最难的一种攻击场景是不允许多次访问的黑盒攻击方法。例如机器学习模型/算法应用于人脸解锁、人脸支付等具体应用时，用户并不能大量地访问目标模型，所以需要基于无访问的黑盒攻击方式攻破目标模型。在这种方式下，攻击者通常会利用对抗样本的迁移性能进行攻击。迁移性是指对一个模型所构造的对抗样本往往也会欺骗其他的黑盒模型。所以攻击者可以不访问目标模型，仅在本地对于构造的类似模型进行攻击，所产生的对抗样本就能够欺骗目标模型。

C.4 对抗攻击的方法

a) 快速梯度攻击是目前被广泛采用的一类攻击方法，其主要思想是寻找深度学习模型的梯度变化最大的方向，按照此方向添加图像扰动，导致模型进行错误的分类。FGSM 以增加对图像分类器损失的方式来对图像添加扰动。通过快速梯度攻击构造对抗样本的优势是效率比较高，最终生成的对抗样本会对原图所有像素点都产生一些微小的扰动。快速梯度攻击作为经典的攻击方式，衍生出了许多以此为基础的攻击方法；
b) 迭代攻击：快速梯度攻击只沿着梯度增加的方向添加 1 步扰动，而迭代攻击则通过迭代的方式，沿着梯度增加的方向进行多步小扰动，并且在每一小步后，重新计算梯度方向，相比 快速梯度攻击能构造出更加精准的扰动，但代价是增大了计算量，同时容易过拟合；
c) 动量迭代攻击：为了更好地平衡攻击性能和泛化能力，动量驱动的迭代攻击是一类典型的方法，在每轮迭代中引入了动量机制，减轻了原本迭代中会产生的震荡现象，并有助于算法逃离局部最优值，从而收敛于全局最优或者更好的局部最优值。

C.5 防御措施

针对上述机器学习模型/算法的安全威胁，可以加强机器学习模型/算法训练与测试，使其构造对抗样本的难度加大，从而使得机器学习模型/算法更加健壮。具体地说，可以尝试以下方法：

a) 对抗训练。对抗训练可通过将各种已知攻击方法生成的对抗样本重新加入训练集，进行重训练，在训练时加入对抗样本的方式仿真测试时可能的数据分本，降低模型对对抗样本识别的错误率，使得最终模型能够抵抗对抗样本攻击；
b) 防御蒸馏。针对机器学习模型、算法的安全威胁，学术界有学者提出如防御蒸馏、对抗训练、输入重构等防御技术对抗闪避攻击。防御蒸馏通过对多个深度神经网络进行串联，将前一个深度神经网络生成的分类结果被用于训练后一个深度神经网络，降低机器学习模型对输入扰动的敏感度，提高机器学习模型稳定性；
c) 输入重构方法。则是通过对输入样本加噪声、去噪声等方法进行变形，这种变形不会影响模型正常分类，但能够一定程度上抵抗对抗样本；
d) 强监督学习方法。传统的监督学习算法通常为端到端的学习，例如训练神经网络时是通过最小化网络预测整体（非线性变换部分+线性分类器部分）的交叉熵。强监督学习方法在训练过程中对非线性变换部分所学的特征施加约束，使得学到的特征对于对抗攻击具有更好的鲁棒性，并且在和线性分类器部分结合之后不影响正常样本上的分类准确率；
e) 对抗样本检测。通过对抗样本和真实样本之前数据分布不一致的特性，构建对抗样本检测器，用于区分正常样本和对抗样本。学术界已经提出了在输入域上进行对抗样本检测，例如图像空间，特征空间，梯度空间等。通常的，由于对抗样本检测器的误判率非常高，所以真正防御的时候经常会和去噪或者重构等预处理方法相结合一起使用。