人工智能安全测评白皮书

（2021）

国家语音及图像识别产品质量监督检验中心

国家工业信息安全发展研究中心人工智能所

2021 年 10 月

编写单位

国家工业信息安全发展研究中心

华为技术有限公司

北京瑞莱智慧科技有限公司

绿盟科技集团股份有限公司

蚂蚁科技集团股份有限公司

北京百度网讯科技有限公司

中国科学院信息工程研究所

北京奇虎科技有限公司

思必驰科技股份有限公司

北京的卢深视科技有限公司

编写人员

朱倩倩、李凯、严敏瑞、舒驰、张旭东、顾杜娟、薛峰、秦栋、郭建领、马多贺、刘昭、赵燕燕、陈智超、贾林、张德岳、唐志敏

前 言

人工智能是引领未来的战略性技术，是带动产业转型升级、推动经济高质量发展的重要引擎，正在引发经济社会和生产生活的深刻变革。但随着人工智能技术在各个领域的广泛应用，其安全问题也愈发突出。为贯彻落实国务院《新一代人工智能发展规划》在人工智能安全方面的部署和要求，提升我国人工智能安全保障能力，支撑人工智能技术与实体经济安全融合，国家工业信息安全发展研究中心联合华为技术有限公司、北京瑞莱智慧科技有限公司、绿盟科技集团股份有限公司、蚂蚁科技集团股份有限公司、北京百度网讯科技有限公司、中国科学院信息工程研究所、北京奇虎科技有限公司、思必驰科技股份有限公司、北京的卢深视科技有限公司共同编制人工智能安全测评白皮书。

本白皮书梳理人工智能安全背景、趋势和范围，分析人工智能安全与可信赖的关系，构建由安全目标、安全风险、安全测评和安全保障为指引的人工智能安全框架，介绍现有人工智能安全测评标准、机构和工具情况，并提出人工智能安全测评建议。从技术和行业的角度以实践案例的形式展示应对人工智能安全风险的解决方案，以供相关人工智能技术、产品、服务提供方和应用方参考，提升自身安全风险防控能力，助力人工智能产业健康发展。

目录

一、人工智能安全内涵

（一） 人工智能安全背景

目前国际上对人工智能的相关定义已逐渐清晰和明确，国际标准ISO/IEC 22989《人工智能概念和术语》给出人工智能分别作为系统和学科的定义。人工智能系统是一组方法或自动化实体，它们共同构建、优化和应用模型以便系统可以针对一组给定的预定义任务计算预测、建议或决定。人工智能学科是研究与人工相关的理论、机制、发展和应用智能。结合现有资料总结，人工智能是能够获取、处理、创建和应用知识，模拟人类感知、学习和决策，以技术、系统的形式完成一个或多个既定任务的理论、方法、技术和应用。

人工智能安全在国际标准中有两层含义，一个对应“safety”，强调功能安全，指免于不能容忍的安全风险，如受控设备和控制系统相关的整体安全；另一个对应“security”，强调信息安全，指除了信息的保密性、完整性和可用性外，也强调其他属性，如真实性、可问责性、不可否认性和可靠性等。

人工智能技术在计算机视觉、语音识别、自然语言处理等方面已经取得了巨大进展，在自动驾驶、智慧金融、智能医疗等重点民生领域深入应用，人工智能系统的安全

性问题频发，对社会稳定、公共安全，甚至是国际政治都可能产生极大的影响。2019 年 12 月，美国加利福尼亚州一辆特斯拉 Model S 在自动驾驶过程中闯红灯发生碰撞事故，造成两名乘客当场死亡。2020 年 4 月，日本东京一辆特斯拉 ModelX 在开启自动驾驶辅助系统 Autopilot 模式后撞上路旁的行人，导致一名男性当场死亡。由于自动驾驶、智慧金融、智能医疗等场景对人工智能的安全、可靠、可控有极高的要求，频发的安全事件引发了大众对人工智能安全的担忧。

（二） 人工智能安全趋势

人工智能安全问题已经成为制约其快速发展和深度应用的重要因素，因此国内外提出发展和安全并重、加强监管和测评的核心思想，规范和保障人工智能的安全发展。

美国依据《2019 财年国防授权法》成立了人工智能国家安全委员会，将人工智能上升到国家安全层级，从顶层统筹指引全美人工智能领域发展，进一步完善人工智能监管链条，为聚合各方之力加速推进人工智能发展奠定重要基础。

欧盟发布《人工智能白皮书——欧洲追求卓越和信任的策略》、《欧洲议会和理事会关于制定人工智能统一规则（人工智能法）和修订某些欧盟立法的条例》等文件，

提出未来人工智能将由强监管转向发展和监管并重，对人工智能系统实行分类分级监管的思想和合规评估要求，旨在欧盟范围上规范和保障人工智能的可信应用。

我国发布《新一代人工智能发展规划》、《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》等多项政策文件，提出加快人工智能安全技术创新，强调安全监管与评估，人工智能安全已成为未来国家战略、政府工作的重要组成部分。

（三） 人工智能安全范围

人工智能作为一项新技术，既会伴生安全问题，又会赋能安全问题。人工智能的伴生安全问题指内生和衍生安全问题。内生安全问题是由于人工智能自身在鲁棒性、可解释性等方面存在的安全隐患或问题；衍生安全问题是应用人工智能技术的系统产生新的安全威胁，攻击者可利用对抗样本或数据投毒技术，对智能系统开展攻击造成功能失效。人工智能的赋能问题指传统安全问题借助人工智能技术得到解决或人工智能技术加剧传统攻击的效率，人工智能可辅助网络空间安全从被动防御趋向主动防御，从而更快更好地识别威胁、缩短响应时间，也可提升网络攻击能力、自动检测网络安全防御方法、制定智能化的攻击策略。

本白皮书中人工智能安全特指伴生安全，指人工智能系统在设计开发、测试验证、部署应用、运行监测、重新评估、退役下线全生命周期的智能水平、可靠性、安全性等风险，即人工智能技术（数据、算法模型）和人工智能技术依赖系统（算力设施、学习框架、云平台）的内生风险和人工智能系统（不当使用、外部攻击、业务设计）的应用风险。

（四） 安全与可信赖的关系

人工智能可信赖包含了安全和伦理两部分内容，人工智能可信赖包含人工智能安全。没有安全人工智能就没有可信赖人工智能，因此保障人工智能安全是实现可信赖人工智能的必由之路。

人工智能可信赖是指满足利益相关方期望并可验证的能力，依赖于具体的产品或服务、数据以及所用技术，应用不同的可信性特征并对其进行验证，以确保利益相关方的期望能得到满足。国家标准《人工智能 术语》报批稿认为可信赖包括服务、产品、技术、数据和信息的可靠性、可用性、韧性、安全性、隐私性、责任性、透明性、完整性、真实性、质量、实用性等，与国际标准ISO/IEC 22989《人工智能概念和术语》中的表述基本一致。

本白皮书中的安全参考在研国标《机器学习算法安全

评估规范》，指人工智能系统在保证符合应用场景功能、性能的前提下，至少应符合保密性、完整性、可用性、可控性、鲁棒性、隐私性等要求。**保密性**是指数据、模型的参数或功能、软硬件依赖信息不可被窃取或非授权访问。完整性是指数据、算法模型、基础设施和产品不被恶意植入、篡改、替换和伪造。**可用性**是指算法可用程度、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。**可控性**是指应在系统、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。**鲁棒性**是指在面对多变复杂的实际应用场景的时候具有较强的稳定性，同时能够抵御复杂的环境条件和非正常的恶意干扰。**隐私性**是指人工智能技术在正常构建使用的过程中，能够保护数据主体的数据隐私，是指在数据信息没有发生直接泄露的情况下，不会间接暴露用户数据。

二、人工智能安全框架

（一）框架总述

人工智能安全框架包含安全目标、安全风险、安全测评、安全保障四大维度，如图 1 所示。人工智能安全实践之路需要四步走，第一步设立人视角的应用安全和系统视

角的技术安全目标，第二步梳理人工智能衍生和内生安全风险，第三步测评数据、算法、基础设施和系统应用风险程度，第四步运用管理和技术相结合的方式保障安全。

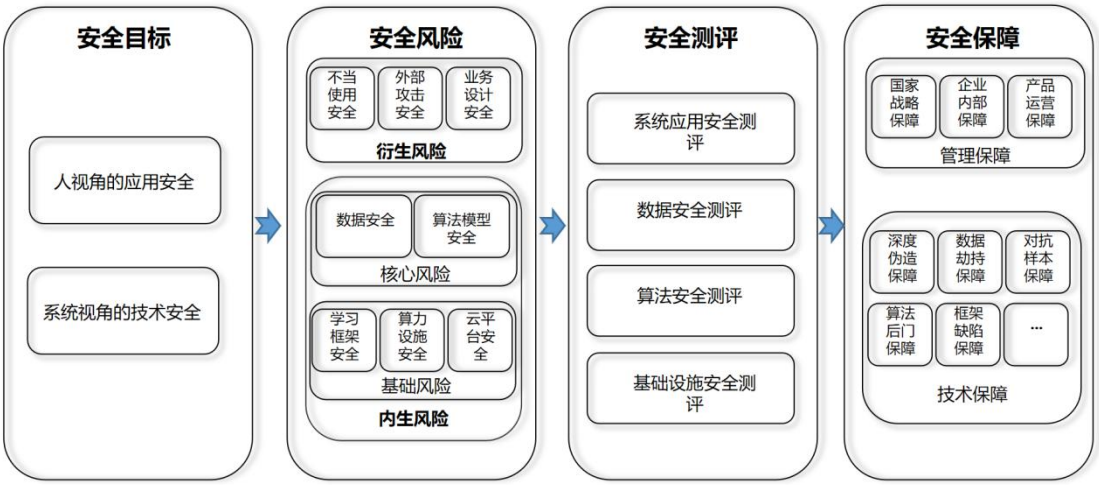

图 1 人工智能安全框架

（二）框架分析

1. 安全目标

国际标准化组织(ISO)和国际电工委员会(IEC)联合发布《信息技术人工智能人工智能的可信度概述》报告中提到由于人工智能技术仍处于快速发展阶段，如果在设计、开发、部署、应用人工智能系统中缺乏最佳实践，则难以通过学习最佳实践提高人工智能系统安全。因此，我们以人的视角和系统的视角提出应用和技术安全目标。

1) 人视角的应用安全

产品功能和目标的设计为系统设计和实施提供参考，

是人工智能产品开发的基础性步骤，这一步骤的失误会严重损害产品呈现的最终效果。为了产品开发需要，应详细、准确考虑到各种情况。例如在自动驾驶场景，要充分考虑到功能和预期功能的安全事件保障和溯源措施。不当使用和外部攻击人工智能系统会给国家、社会、企业和个人带来严重危害。因此，应确保人工智能系统在设计功能时准确全面的考虑场景和突发状况，对不当使用进行适度的监管控制，对应已知和未知的外部攻击提供相应的保障机制，并符合国家法律法规和行业管理规范的要求。

2) 系统视角的技术安全

在人工智能系统开发、验证、部署、维护、退役的全生命周期，都存在数据、算法模型、基础设施的保密性、完整性、可控性风险。特别在公共服务、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安全要求应更为严格。因此，在人工智能系统在生产实施时，应确保数据采集、传输、存储、处理、交换、销毁过程的安全，算法模型设计开发、验证测试、部署运行、维护升级、退役下线过程的安全，承载系统的学习框架、算力设施、云平台安全。

2. 安全风险

随着人工智能产业的快速发展，算法黑箱、技术滥用、侵犯隐私等安全问题逐步显现，给公共安全、道德伦理、社会治理等带来挑战。主要的人工智能安全风险可以归纳为技术内生风险和系统衍生风险，人工智能安全风险如图 2 所示。

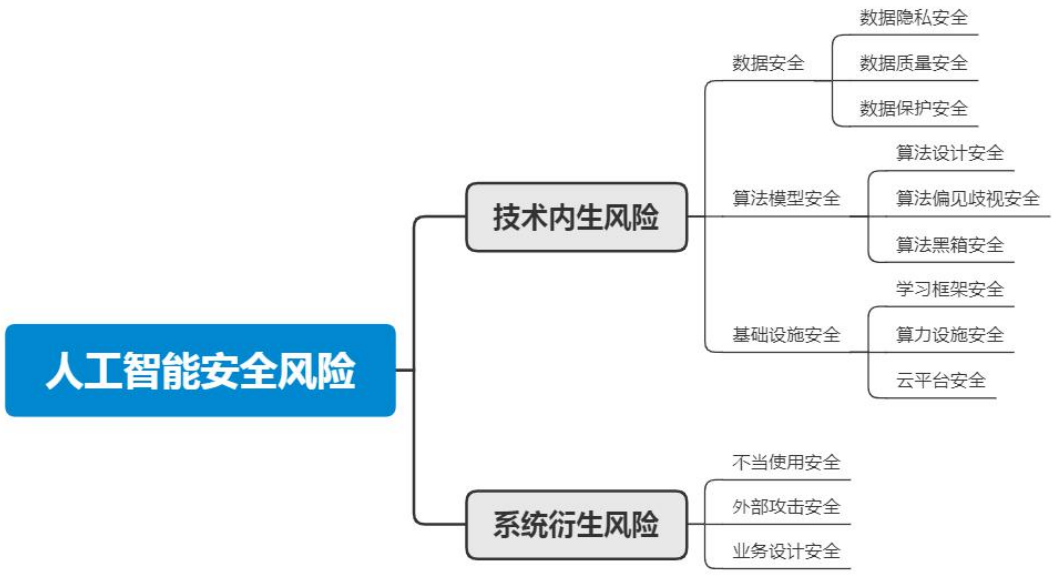

图 2 人工智能安全风险

1) 技术内生风险

数据是人工智能的重要驱动，算法模型是人工智能的核心技术，基础设施是人工智能的底层支撑，技术内生风险为由于数据安全风险、算法安全风险和基础设施安全风险对社会安全产生重大威胁。

数据安全风险。数据安全风险包括数据隐私、数据质

量、数据保护安全风险。数据隐私安全风险是人工智能的开发、测试、运行过程中存在的隐私侵犯问题；数据质量安全风险是用于人工智能的训练数据集以及采集的现场数据潜在存在的质量问题，以及可能导致的后果；数据保护安全风险是人工智能开发及应用企业对持有数据的全生命周期安全保护问题。2020 年 2 月，服务于 600 多家执法机构及安保公司的美国人脸识别创业公司 ClearviewAI 称其客户面部信息数据库被盗，据报道，ClearviewAI 从网络社交媒体上抓取了超过 30 亿张照片，这些数据在采集时并未明确获得用户的同意，涉嫌侵害个人信息泄露。

算法安全风险。算法安全风险包括算法设计、算法黑箱和算法偏见歧视风险。算法设计安全风险是在算法或实施过程有误可产生与预期不符甚至伤害性结果；算法黑箱安全风险是基于神经网络的深度学习算法，通过复杂的计算过程对输入数据进行计算，人们依据现有的科学知识和原理对输出的结果难以解释；算法偏见歧视风险是在信息生产和分发过程失去客观中立的立场，影响公众对信息的客观全面认知，或者在智能决策中，通过排序、分类、关联和过滤产生不公平问题，主要表现为价格歧视、性别歧视、种族歧视。

基础设施风险。基础设施风险包括学习框架、算力设施和云平台安全风险。云端框架安全风险主要来自于自身

代码的实现以及第三方的依赖库问题；终端框架安全风险主要存在于模型网络参数、模型网络结构，以及模型转换过程。算力设施安全风险主要是 GPU 驱动和芯片漏洞问题。云平台安全风险主要是虚拟化和 Web 平台问题，用于深度学习任务的节点性能强大，会有一些攻击者想要非法使用这些资源进行挖矿。

2) 系统衍生风险

技术本身并无对错之分，但不当使用、外部攻击、业务设计问题会给国家安全、社会安全、人身安全带来挑战。

不当使用安全风险。人脸识别技术能有效维护社会治安，也能给日常生活提供便利，而滥用、盗用、伪造个人生物特征却给社会治安带来严重危害。2021 年央视 3.15 晚会曝光了科勒卫浴等知名品牌安装人脸识别摄像头，捕捉顾客人脸信息，分析顾客到访次数来区别对待,严重侵害了消费者隐私。

外部攻击安全风险。由于现实应用场景的开放性，人工智能系统面临对抗样本、模型窃取等恶意攻击，人工智能安全性和可信赖性受到质疑。2019 年，比利时鲁汶大学研究人员在胸前粘贴一张设计好的打印出来的对抗图案，成功避开 AI 视频监控系统，自由出入门禁。2021 年，瑞莱

智慧公司的研究人员在一项测试中，通过佩戴一副含有对抗样本图案的眼镜，破解了 19 款常见安卓手机的人脸识别解锁系统。

业务设计安全风险。主要指功能安全，包括预期功能安全，分为应用部署和运行维护两个方面。在应用部署时，应充分考虑对周边环境的影响，防止危险故障的发生或在出现故障时能够进行有效控制，尽可能降低事故发生时造成的附带财产以及人员损失；在运行维护时，要求设备在设计时具有在人机交互层面对于安全隐患的考虑，全面识别受控设备中的所有危险。2018 年，美国亚马逊公司机器人误戳穿驱熊喷剂罐，导致 24 名工人受伤。

3. 安全测评

安全测评指依据法律法规、行业规范和标准指南，对人工智能系统的基础设施、核心技术和业务应用在一定的安全机制下通过检查和测试相结合的方式进行评估。通过企业自测评和三方监管测评保障技术安全、可靠、可控，为我国人工智能产业持续健康发展提供支撑。测评的范围主要包括数据安全测评、算法安全测评、基础设施安全测评、系统应用安全测评。

1) 数据安全测评

数据安全是组织机构信息安全体系的重要环节，通过统一标准《数据安全能力成熟度模型》来评估机构数据安全能力，用组织的能力成熟度来进行安全风险评估，技术维度对数据生存周期安全过程维度进行测评，包括数据采集安全、数据传输安全、数据存储安全、数据处理安全、数据交换安全、数据销毁安全。管理维度对组织个数据安全过程应具备安全能力的评估，包括组织建设、制度流程、技术工具、人员能力，如图 3 所示。通过评估发现数据安全能力短板，查漏补缺，提升行业的整体安全管理水平和产业竞争力。

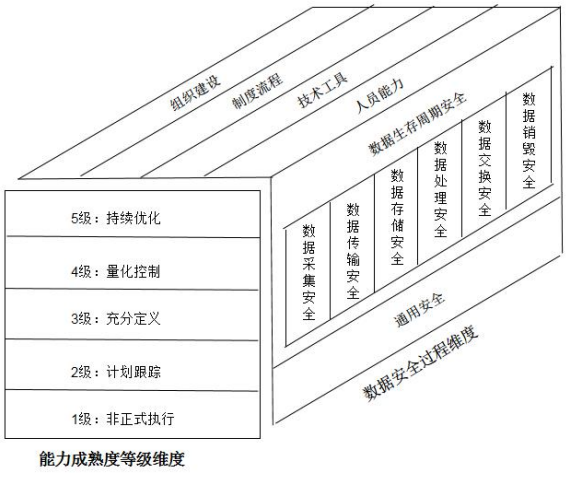

图 3 数据安全能力成熟度模型架构图

2) 算法安全测评

算法安全是人工智能系统应用安全的核心要求，结合国内应用实践和标准编制组的研究成果，提出与国际标准

接轨、适合我国国情的算法安全标准《机器学习算法安全评估规范》、《信息技术人工智能机器学习模型及系统的质量要素和测试方法》、《信息安全技术人脸比对模型安全技术规范》，形成了机器学习算法和模型的通用质量要素和测评方法，规定了人脸比对模型在应对对抗样本方面的技术要求、测试方法和安全分级，覆盖人工智能算法全生命周期的安全测评指标和测评体系，提升机器学习算法实践应用的安全保障能力。

3) 基础设施安全测评

基础设施安全是人工智能系统建立的根基保障，人工智能关键基础设施安全问题易被忽视，如学习框架自身代码实现、第三方图象处理库、算子库、GPU 驱动、芯片安全、虚拟化、web 平台等安全风险。可参考国际标准《深度学习软件框架评估方法》、国家标准《服务器安全技术要求和测评准则》、联盟标准《绿色计算服务器可信赖技术要求》，对学习框架和算力设施的安全稳定性评估，降低人工智能基础设施遭受的攻击风险，增强客户对绿色计算体系服务器的信任度，助力绿色计算服务器加速产业应用。

4) 系统应用安全测评

系统应用安全是人工智能安全最佳的实践验证，人工智能系统应用涉及多个行业和领域，覆盖多项技术协同实施，呈现多种智能产品服务形式。需针对技术、产品、服务和行业领域特点，依据法律法规、行业规范、标准指南进行符合性评估，依据产品设计文档对人工智能系统进行功能和预期功能的使用范围、潜在受影响的人员数量、应对外部攻击能力、对结果依赖性和损害的不可逆转性等风险性评估。

4. 安全保障

人工智能安全保障体系中，政府发挥着领导性作用，企业发挥着决定性作用，政府和企业需共同努力在管理层和技术层保障人工智能系统的安全，管理层分为战略保障、企业保障和运营保障。

1) 战略保障

在宽松政策和内部需求的驱动下，人工智能应用成果将会加速落地，一系列安全问题的暴露促使我国建立越来越完善的国家战略，主要包括法律法规、行业规范、标准指南，护航人工智能产业的安全发展，实现产业智能化升级的目标。

a. 法律法规

近年来，我国加快了人工智能的立法进程，在数据安全、算法安全、基础设施安全、应用安全方面都有涉及，在数据安全和个人信息保护方面的力度较大。

在数据安全方面，2017 年《网络安全法》防止网络数据泄露或被窃取、篡改，采取数据分类、重要数据备份和加密等措施。2021 年第十三届全国人大常委会第二十八次会议对《个人信息保护法（草案二次审议稿）》进行了审议，规定敏感个人信息、国家机关处理个人信息、个人信息跨境的规则，保护个人信息权益,规范个人信息处理活动,保障个人信息依法有序自由流动,促进个人信息合理利用。2021 年《数据安全法》规范数据处理活动，保障数据安全，促进数据开发利用，保护个人、组织的合法权益，维护国家主权、安全和发展利益，特别对行业组织提出了制定安全行为规范，加强行业自律，指导会员加强数据安全保护的要求。这项法规有效的消灭了灰色地带，对各行业都形成了法律约束，杜绝了数据的随意共享和流转。

在算法安全方面，2021 年国家网信办发布《互联网信息服务算法推荐管理规定（征求意见稿）》。征求意见稿中提到，参与算法推荐服务安全评估和监督检查的相关机

构和人员应当对在履行职责中知悉的个人信息、隐私和商业秘密严格保密，不得泄露、出售或者非法向他人提供。

在基础设施安全方面，2021 年《关键信息基础设施安全保护条例》是《网络安全法》的实施细则，依照本条例和有关法律、行政法规的规定以及国家标准的强制性要求，在网络安全等级保护的基础上，采取技术保护措施和其他必要措施，应对网络安全事件，防范网络攻击和违法犯罪活动，保障关键信息基础设施安全稳定运行，维护数据的完整性、保密性和可用性。

在应用安全方面，2020 年中国十三届全国人大三次会议表决通过了《民法典》规定任何组织或者个人不得以丑化、污损，或者利用信息技术手段伪造等方式侵害他人的肖像权和声音。《App 违法违规收集使用个人信息行为认定方法》、《常见类型移动互联网应用程序必要个人信息范围规定》为监督管理部门认定 App 违法违规收集使用个人信息行为提供了参考同时也为 App 运营者自查自纠和网民社会监督提供了指引。2021 年《道路交通安全法（修订建议稿）》，将自动驾驶交通安全纳入管理范畴，制定了测试、上路、事故责任等相关规定内容，内容虽简短，但对于推动智能网联汽车商业化应用具有重要保障意义。

b. 行业规范

我国配套法律法规的人工智能行业规范的正在逐步跟上，对于自动驾驶、智慧金融、智能医疗等人工智能应用主要以主管部门的政策指导实施监管和规范。

在自动驾驶领域，2018 年《智能网联汽车道路测试管理规范（试行）》依据《道路交通安全法》《公路法》等法律法规制定，规范智能网联汽车道路测试管理。2021 年《智能网联汽车道路测试与示范应用管理规范（试行）》适应行业新的发展需求，推动实现由道路测试向示范应用扩展。同年《关于促进道路交通自动驾驶技术发展和应用的指导意见》、《关于加强智能网联汽车生产企业及产品准入管理的意见》要求加强汽车数据安全、网络安全、软件升级、功能安全和预期功能安全管理。同年《汽车数据安全管理若干规定（试行）》在汽车数据安全管理领域出台有针对性的规章制度，明确汽车数据处理者的责任和义务，规范汽车数据处理活动。

在智慧金融领域，2020 年《个人金融信息保护技术规范》规定了个人金融信息在收集、传输、存储、使用、删除、销毁等生命周期各环节的安全防护要求，从安全技术和安全管理两个方面，对个人金融信息保护提出了规范性要求。同年《金融数据安全分级指南》给出了金融数据安

全分级的目标、原则和范围，明确了数据安全定级的要素、规则和定级过程，并给出了金融业机构典型数据定级规则。

在智能医疗领域，2017 年《人工智能辅助诊断技术管理规范》《人工智能辅助治疗技术管理规范》分别对使用计算机辅助诊断软件及临床决策支持系统和使用机器人手术系统辅助实施手术的技术提出了规范要求。2018 年《全国医院信息化建设标准与规范（试行）》利用人工智能技术对疾病风险进行预测，实现医学影像辅助诊断、临床辅助诊疗、智能健康管理、医院智能管理和虚拟助理。

c. 标准指南

在《国家新一代人工智能标准体系建设指南》的引领下，人工智能国际标准、国家标准、行业标准、团体标准等的制修订和协调配套，在指南的指导下结合产业界实际情况，形成人工智能安全标准体系框架，如图 4 所示，相关标准的梳理，如图 5 所示。

图 4 人工智能安全标准体系框架

基础安全标准：《人工智能人工智能可信度概述》、《机器学习模型与人工智能系统的可解释性方法及目标》规范人工智能安全相关基本概念，参考架构等。

计算设施安全标准：《服务器安全技术要求和测评准则》、《绿色计算服务器可信赖技术要求》、《计算硬件在人工智能安全领域的角色》对人工智能计算设施有参考意义，《服务器机密计算参考架构及通用要求》、《基于安全多方计算的隐私保护技术指南》、《多方安全技术金融应用技术规范》、《共享学习系统技术要求》指导人工智能应用/系统/产品等提供开发、运行环境的算力资源安全、支撑工具安全等，具体可包括人工服务器安全、学习框架安全等方面。

数据安全标准：《人脸识别数据安全要求》、《步态识别数据安全要求》、《基因识别数据安全要求》、《声纹识别数据安全要求》规范人工智能应用/系统/产品等开

发运行过程中对数据及个人隐私的安全保护，主要关注人工智能数据集相关标准，以及个人信息保护等。

算法和模型安全标准：《机器学习算法安全评估规范》规范机器学习算法模型全生命周期的安全要求和证实方法，以及机器学习算法模型的安全评估实施。

应用和产品安全标准：《生物特征识别信息的保护要求》、《远程人脸识别系统技术要求》、《虹膜识别系统技术要求》、《指纹识别系统技术要求》、《智能家居安全通用技术要求和测试评价方法》、《智能门锁安全技术要求和测试评价方法》、《智能终端设备安全能力技术要求》、《智能家居网络系统安全技术要求》、《移动智能终端人脸识别安全技术要求及测试评估方法》、《人工智能系统测试》、《安防人脸识别应用防假体攻击测试方法》、《生物特征识别防伪技术要求第 1 部分：人脸识别》、《基于可信环境的生物特征识别身份鉴别协议》规范人工智能相关应用和产品自身安全。

管理和服务安全标准：《人工智能风险管理》、《人工智能对隐私的影响》、《可信人工智能研发管理指南》、《ICT 供应链安全风险管理指南》、《人工智能风险评估模型》规范人工智能系统/应用/产品安全、稳定运行的一些管理和服务安全，包括风险评估、研发过程管理等。

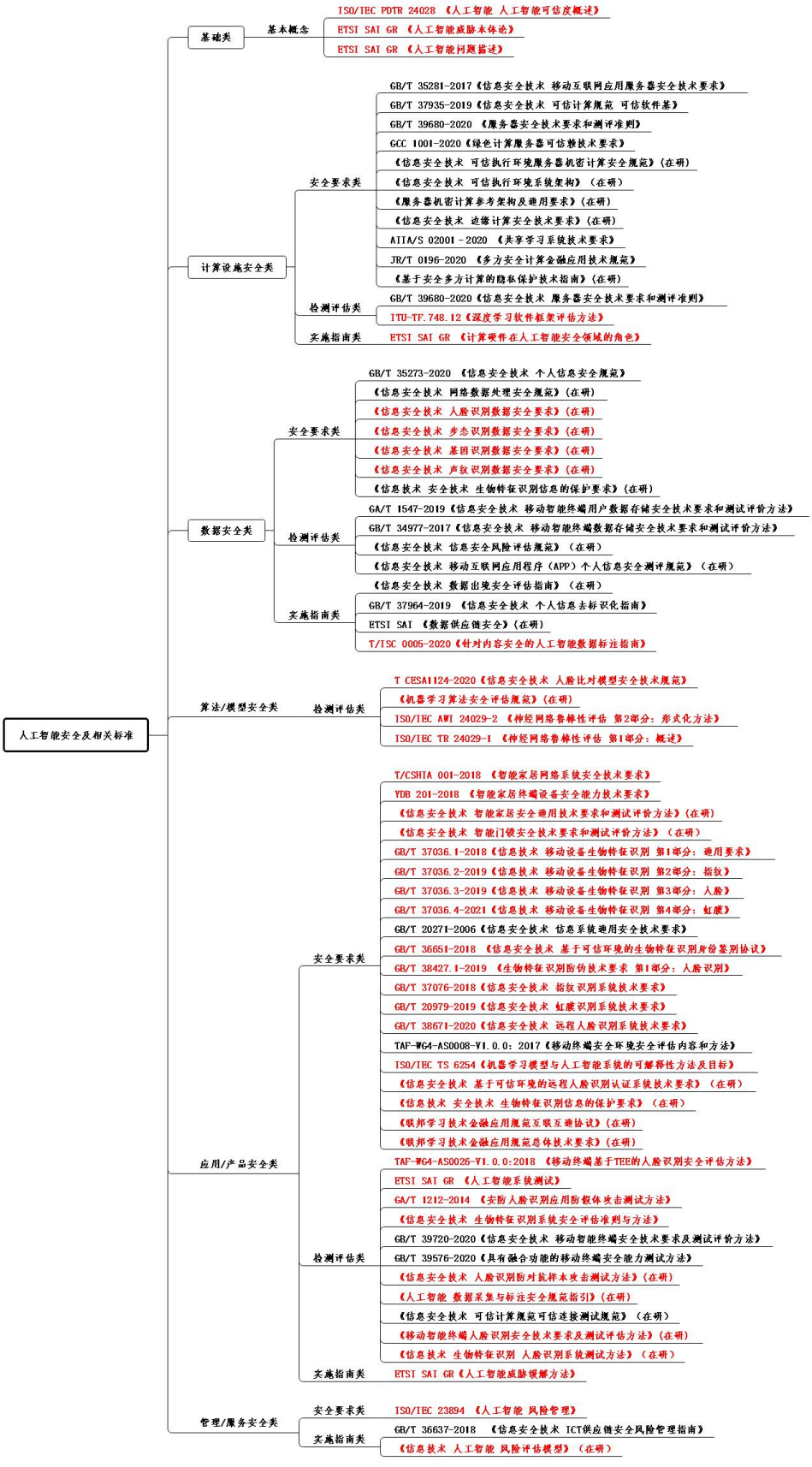

图 5 人工智能安全及相关标准

2) 企业保障

在我国多项法律法规中有机制保障制度，国家不通过检查的方式进行实施，而是监管的方式进行事后处罚。因此企业内部的管理要求显得至关重要。企业保障体系主要包括组织设计、人员管理和流程管理，如图 6 所示。

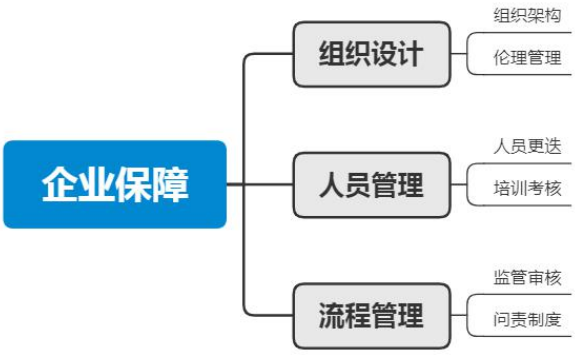

图 6 人工智能企业保障体系

a. 组织设计

企业设立独立的人工智能安全管理部门，配备足量的专业的管理人员，定期对战略层法律法规、行业规范、标准指南等文件进行解读分析，跟踪人工智能安全技术和应用变化，优化人工智能安全管理要求和应急处理方案，制定符合业务人员操作的人工智能安全要求，实现操作流程的及时更新。设立独立的人工智能安全监督部门，安排专人对企业要求的落实情况进行评估和汇报。对人工智能安全相关的业务人员在人员管理、业务功能、机制设定、伦理合规和成本效率等方面进行培训，落实岗位分工和岗位

职责。

b. 人员管理

人才是确保任何一个企业具备竞争力的核心因素，因此应具备专门的人员更迭管理团队，对关键岗位和对应业务人员的更迭情况有自动化、数字化的记录管理，对人员更迭中的异常情况进行实时监控，并进行追访回溯。人工智能安全相关人员主要包括承担决策责任的管理人员、参与开发设计的技术人员和负责运维的运营人员。应对以上人员进行适当的培训和考核，确保满足各自的岗位需求，并对考核中发现的问题进行汇总，及时更新培训考核机制。

c. 流程管理

具备对人工智能系统的监控、响应、升级处理与审核流程，建立完全独立的运营体系，运营数字化、自动化、实时化，具备自我学习优化闭环能力，对监控与审核的过程有完备的存档体系，安全存储，不可篡改。具有明确的问责制度，对相关人员有清晰、公开的责任划分，包括对各环节涉责管理人员有必要的责任追溯，对重要的事件，核心管理人员有必要的连带责任追溯。

3) 运营保障

运营体系的构建可以为人工智能系统的应用提供循环优化保障，通过事前风险评估、事中监测预警、事后安全审计，支撑安全运营全生命周期智能化发展。

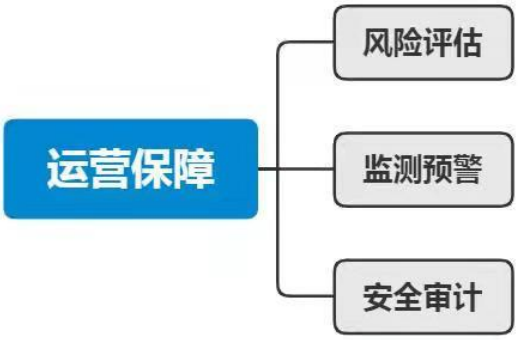

图 7 人工智能运营保障体系

a. 风险评估

建立贯穿高风险人工智能系统生命周期的风险管理系统，识别、分析和评估人工智能系统风险并采取相应的风险管理措施，通过设计开发、使用者培训等方法降低风险。高风险人工智能系统的供应商建立质量管理体系，做好系统设计、开发等过程的质量管理和记录。编制高风险人工智能系统的技术文档，文档内容涉及系统总体描述、系统具体要素和研发过程等等，技术文档应随系统的使用进行持续更新，保持最新状态。高风险人工智能系统的设计、开发和使用需配备适当的人机界面工具，以便在使用期间可由自然人进行有效监督、干预系统运行等，发挥人

工监督的兜底作用。

b. 监测预警

自动记录人工智能系统运行情况日志，日志应具备系统全生命周期的可追溯性以便于监测运行，记录内容包括系统使用时间、系统决策参考的数据库、产出结果的输入数据、特殊系统产出结果的鉴定人、基础设施的运行状态数据和产生的网络数据等。对日志中的数据进行分析,监测关键安全性能,并预测安全性能的变化趋势，根据安全性能的变化趋势生成相应的预警信息。

c. 安全审计

一旦出现安全问题时，企业应对安全问题开展审计。审计内容主要包括风险管理、数据管理、技术文档、记录保持、对用户透明和信息提供、人工监督，以及准确性、鲁棒性和安全性等方面是否符合立法规定的要求，并对发现的问题进行风险评估和整改，形成良性循环。

4) 技术保障

技术内在安全是人工智能安全的核心要素，针对第二代人工智能技术自身的缺陷，需通过技术手段检测和防御加固模型，应对深度伪造、对抗样本、算法后门、数据劫

持、框架缺陷通用新型人工智能安全风险。

a. 深度伪造

深度伪造技术通过机器学习模型将图片或视频合并叠加到源图片或视频上，可以将个人的声音、面部表情及身体动作拼接合成虚假内容。

检测方法：目前深度伪造检测技术有两种。一种是基于特定特征的方法，深度伪造内容通常会产生特定特征，这类特征可以使用取证分析以及机器学习方法来检测。分为以下七个类别：混合空间特征、环境空间特征、取证空间特征、行为时序特征、物理时序特征、同步时序特征、连贯时序特征。另一种是基于非特定特征的方法，相较于检测特定特征，基于非特定特征的方法训练神经网络作为通用分类器，让网络决定要分析哪些特征。一般来说，基于非特定特征的检测方法分为以下两种：分类或异常检测。

防御方法：目前深度伪造防御方法主要有两种。一种是采用联合深度学习方法，通过结合矩阵盲化技术和随机梯度下降法，可以实现输入向量以及部分模型参数的盲化。通过限制攻击者本地生成对抗网络的建模与更新，同时限制深度学习模型使用权等方式，允许分布式训练者在本地利用隐私数据集训练得到模型参数的梯度更新，每个

训练者的梯度更新将由参数服务器聚合，实现系统模型的全局更新。另一种是使用区块链技术，通过利用区块链技术，实现对图像、音频和视频的真实性证明，从而防御深度伪造内容，例如使用以太坊智能合约，利用用于存储数字内容及其元数据的星际文件系统(IPFS)的哈希值，即使数字内容被多次复制，仍可以追踪和跟踪到数字内容的出处及其原始来源。

b. 对抗样本

对抗样本攻击是利用对抗样本对机器学习算法进行欺骗的恶意行为。对抗样本是指对正常数据集通过故意添加干扰所形成的恶意输入样本，这种恶意样本可以导致机器学习算法出现指定的错误预测结果。

检测方法：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。学术界已经提出了多个有代表性的对抗样本检测方法，例如可以从从训练数据的流形（manifold）的角度去理解对抗样本和真实样本的差异，通过核密度估计的方法进行对抗样本检测；也可以通过提取神经网络不同层输出的局部本征维数，进而运用分类器进行对抗样本样本检测等。

防御方法：目前对抗样本的防御技术中最主流的方法

是对抗训练技术，对抗训练通过将模型动态生成的对抗样本以数据增强的方式引入到训练过程中，使得模型学习对抗样本中的特征，从而提升模型对于对抗样本的防御能力。除直接通过对抗训练或可认证防御等技术提升模型鲁棒性外，还可以通过梯度梯度隐藏或模型随机化等方法使攻击者无法获得准确的反馈信息，从而增加攻击者构造对抗样本的难度，进而实现防御。另外，在输入模型进行预测之前，先对输入的图像进行重构变化，在保留图像语义的情况下破坏对抗干扰，也能够在一定程度上防范对抗样本攻击。

c. 算法后门

算法后门攻击是一种针对模型训练完整性的威胁，在训练阶段添加敌手选定的后门触发器，向人工智能模型插入后门，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，因此具有很强的隐蔽性。

检测方法：由于遭受后门攻击的算法模型只有面对嵌入后门触发器的输入数据时才会触发恶意行为，因而检测人工智能应用的算法模型是否遭受后门攻击极具挑战性，也成为了当前研究的热点。例如，美国加州大学、芝加哥大学和弗吉尼亚理工大学研究人员联合提出神经元清洁检

测方法，该方法利用算法后门构建了将其他类数据以非常小扰动转变为目标类路径的特点，通过遍历计算所有标签作为目标标签所需要的输入数据最小扰动量来检测算法后门。加州大学圣地亚哥分校首次提出了黑盒场景下的后门检测方案，利用条件生成模型来学习潜在后门触发器的概率分布，从而反向生成模型的后门触发器，对其扰动水平进行检测评估，确定模型异常的阈值。普渡大学提出了 ABS 模型检测方法，通过扫描 DNN 模型并对比 DNN 模型内部个体神经元活性差异来确认模型是否存在后门。

防御方法：现有的防御方法大多针对的是基于投毒的后门攻击，可分为启发式后门防御和可验证后门防御两类。启发式后门防御是基于对现有攻击的观察实验所设计的防御方法，在实践中能达到良好的防御效果，但其有效性缺乏理论支撑；而可验证后门防御的有效性会有一定的理论支撑，但在防御效果上通常不如启发式防御方案。启发式后门防御包括了基于预处理的防御、基于模型重构的防御等方式。基于预处理的后门防御采用不同的预处理手段对样本进行处理，例如，使用预训练的自动编码器作为预处理器、利用风格迁移对图像做预处理、利用空间变幻对输入做预处理等，预处理的作用在于抑制后门的激活。基于模型重构的防御方法目标是去除毒化模型中隐藏的后门，从而在面对带有触发器的输入时，模型的预测不会出

现被影响。

d. 数据劫持

数据劫持指在数据采集、传输、使用过程中，通过对设备硬件、操作系统、应用程序或网络链路，使用注入、替换等技术拦截或篡改数据。

检测方法：由于在各个环节都可能发生数据劫持，需要针对不同环节使用不同的检测方法。一是网络劫持检测，需要检查所在的网络是否使用明文、不安全的协议、若加密算法、不安全的网络环境等；二是应用篡改检测，验证应用的完整性、以及运行过程中应用的状态、扫描用户内存是否被异常篡改；三是操作系统检测，检查所处的操作系统是否存在异常模块，风险框架，非标准API，被篡改的系统库等；四是设备硬件检测，检查设备的硬件是否合法，是否被替换，检查传感器等数据是否出现异常等；五是业务流程检测，检查程序运行中是否出现不符合设计预期的调用或用户操作。

防御方法：针对数据劫持攻击，需要对数据的采集到使用的各个环节增加风险检测、数据加密和逻辑校验，实现人工智能应用的全流程安全防护。一是符合商用密码认证要求，使用符合国密要求的密钥算法和相应使用要求，防止攻击者对密文数据进行解密。二是程序的自我保护，

程序运行时拥有独立进程、独立文件存储等保护措施，防止其他程序对自己的非法读取以及修改，增加攻击者逆向和 HOOK 的难度。三是风险检测，采集数据的同时，还需采集设备信息，对危险设备、危险环境、危险行为进行识别。四是传输链路保护，数据传输时，增加全链路的数据校验能力，拦截校验失败的数据，网络传输应使用 HTTPS 协议，保障通讯安全，同时校验 CA 证书，防止中间人攻击，应建立安全 DNS 解析，防止域名劫持。五是业务流程检查，使用数据时，增加对业务流程的检查，拦截不符业务流程规范的调用，防止攻击者绕过风控检测。

e. 框架缺陷

框架缺陷是指机器学习框架代码中存在问题、缺陷，致使学习框架在训练或者推理时出现不满足预期的行为。

检测方法：框架缺陷通常有静态分析和动态分析两种方式。静态分析技术是一种无需运行程序，而对程序进行安全性分析的技术。静态分析通常使用模式匹配、上下文环境搜索等方式查找不正确的函数调用及异常的函数返回状态，挖掘程序中可能存在的缺陷。动态分析技术是指在程序运行期间检测其安全性的技术，常用的动态分析技术包括模糊测试、符号执行等。模糊测试的流程分为三步：第一步，构造畸形数据作为测试用例；第二步，将测试用

例输入给目标程序；第三步，监控目标程序执行过程中是否产生异常行为。

防御方法：为了保障学习框架的安全性，需要从产品开发到发布使用的不同阶段进行安全性防御。一是开发流程引入 DevSecOps，DevSecOps 能够在保证原有研发效率的基础上，引入安全属性，保障整个产品开发生命周期中的健壮性。二是强化关键点防御，开发者使用学习框架进行神经网络的训练、推理时，主要使用框架提供的功能算子，因此需要着重对不同功能算子的安全性防护，增强功能算子的安全性防御能力。三是保障运行环境可靠性，由于云平台的兴起，使得不同用户可在同一个平台中进行模型的训练、推理，因此需要保障学习框架环境的安全性、隔离性，防止由于单例用户的安全问题危害平台其他用户。

三、人工智能安全测评现状

（一）安全测评概述

欧洲联盟、亚太经济合作组织、东南亚国家联盟、海湾国家合作委员会等区域性政府间组织，以及越来越多的国家政府，在相关产品的市场准入规则和监管中，提倡和采信利用获认可的检验检测认证来提高市场监管的效能，促进市场准入便利化。

检验检测行业是高技术服务业、生产性服务业、科技服务业，具有公共保障性和市场开放性的特征，检验检测与计量、标准、认证认可共同构成国家质量基础设施，是现代服务业的重要组成部分。检验检测在服务国家经济发展、服务产业科技发展、保障社会安全、保障人民健康方面发挥着重要的支撑和引领作用。

我国依据国际标准 ISO/IEC 17025《检测和校准实验室能力的通用要求》，强制性实施检验检测机构资质认定制度和自愿性国际互认实验室认可制度，实施模式（程序）大体相同，本质上都是对实验室的检测能力和管理体系是否满足标准要求的一项资质评价制度。

检验检测机构资质认定（CMA）制度是国务院确定的行政许可项目，向社会出具具有证明作用的数据和结果的检验检测机构应当取得资质认定。自 1987 年开始实施至今，目前成为我国认知度最高、影响最大、国务院部门间合作最多、获证机构数量最多的检验检测机构市场准入制度，为各行各业提供了基准统一、通用开放、权威可信的资质评价服务与管理保障。

实验室认可（CNAS）制度是依据标准证实认证机构、实验室等合格评定机构具有特定的技术和管理能力，表明合格评定机构具备特定能力的一种第三方证明。它的核心是认可机构对于满足要求的合格评定机构予以正式承认，

并颁发认可证书，以证明该机构具备实施特定合格评定活动的技术和管理能力。通过认可后，机构的能力得到证实，出具的证书和报告更能被政府管理部门和公众所信任。

认可分为对认证机构、实验室、检验机构的认可。在市场经济中，实验室是为贸易双方提供检测、校准范围的技术组织，实验室需要依靠其完善的组织结构、高效的质量管理和可靠的技术能力为社会与客户提供检测、校准服务。认证认可属于第三方合格评定活动，与其他形式的合格评定活动的最大区别就在于它是由独立、权威和具有较强专业背景的机构所进行的合格评定，并通过书面形式对评定结果加以公示证明。其结果能够更好地获得需求方的信任，从而成为合格评定活动中最重要、最核心的组成部分。

通过梳理第三方测评机构作为测评依据的人工智能安全标准，对现有依据人工智能标准进行测试评估的检验检测认证机构进行不完全统计，分析现有测评机构和安全检测工具。

（二）安全测评标准现状

根据中国电子技术标准化研究院公布数据，截止到2021年7月，国际标准化组织在研和已发布的人工智能相

关标准共 31 个，我国在研、拟研制和已发布人工智能标准 215 个，人工智能安全标准 33 个。经过对人工智能安全相关的测评标准进行梳理和补充，形成第三方测评机构作为测评依据的人工智能安全及相关标准 15 个，如表 1 所示。作为第三方测评机构测试依据的标准主要集中在产品/应用安全类，涉及国家标准、行业标准和团体标准。

表 1 第三方测评机构作为测评依据的人工智能安全标准

序号	类别	标准编号	标准名称
1.	数据安全类	GB/T 34977-2017	《信息安全技术 移动智能终端数据存储安全技术要求和测试评价方法》
2.		GB/T 34978-2017	《信息安全技术 移动智能终端个人信息保护技术要求》
3.	应用产品安全类	GA/T 1212-2014	《安防人脸识别应用防假体攻击测试方法》
4.		GB/T 34975-2017	《信息安全技术 移动智能终端应用软件安全技术要求和测试评价方法》
5.		GB/T 34976-2017	《信息安全技术 移动智能终端操作系统安全技术要求和测试评价方法》
6.		T/CSHIA 001-2018	《智能家居网络系统安全技术要求》
7.		TAF-WG4-AS0026-V1.0.0:2018	《移动终端基于 TEE 的人脸识别安全评估方法》
8.		YDB 201-2018	《智能家居终端设备安全能力技术要求》
9.		GB/T 38671-2020	《信息安全技术 远程人脸识别系统技术要求》
10.		GB/T 38632-2020	《信息安全技术 智能音视频采集设备应用安全要求》
11.		GB/T 39720-2020	《信息安全技术 移动智能终端安全技术要求及测试评价方法》
12.		GB/T 37076-2018	《信息安全技术 指纹识别系统技术要求》
13.		GB/T 20979-2019	《信息安全技术 虹膜识别系统技术要求》
14.		GB/T 37036.3-2019	《信息技术 移动设备生物特征识别 第 3 部分：人脸》
15.		GB/T 38427.1-2019	《生物特征识别防伪技术要求 第 1 部分：人脸识别》

注：红色字体为人工智能安全标准，蓝色字体为人工智能安全相关标准。

数据安全方面，国家标准《信息安全技术 移动智能终端数据存储安全技术要求和测试评价方法》、《信息安全技术 移动智能终端个人信息保护技术要求》规定了移动智

能终端数据存储的安全技术要求、测试评价方法及安全等级划分，移动智能终端的个人信息分类及个人信息的保护原则和技术要求。

产品/应用安全方面，国家标准《信息安全技术 远程人脸识别系统技术要求》、《信息安全技术 指纹识别系统技术要求》、《信息安全技术 虹膜识别系统技术要求》、《信息技术 移动设备生物特征识别 第 3 部分：人脸》、《生物特征识别防伪技术要求 第 1 部分：人脸识别》、《信息安全技术 智能音视频采集设备应用安全要求》、《信息安全技术 移动智能终端安全技术要求及测试评价方法》，行业标准《安防人脸识别应用防假体攻击测试方法》，团体标准《移动终端基于 TEE 的人脸识别安全评估方法》对生物特征识别系统的功能、性能、安全性作出要求，对智能音视频采集设备、移动智能终端等智能产品建立完整的测评指标和测评方法。

（三）安全测评机构现状

1. CMA 认定的检验检测机构

我国现有获得资质认定的检验检测机构 3399 家，其中国家级检验检测机构有 836 家。据不完全统计，在人工智能安全领域具有检测能力的检验检验机构有 44 家，如表 2 所示，其中通过国家级资质认定的检验检测机构有 17 家。

从机构性质上看，主要分布在公安部、工信部等国家部委事业单位，公安体系 9 家，工信体系 11 家，处于国有机构把控。从地域分布上看，主要集中在北京、上海、广东和江苏，人工智能产业发展较好的地区，如图 8 所示。

表 2CMA 认定人工智能安全及相关检验检测机构

序号	证书号	机构名称	场所地址
1.	200009344677	国家语音及图像识别产品质量监督检验中心/国家工业信息安全发展研究中心人工智能所	北京市市辖区石景山区鲁谷路 35 号
2.	170009013402	中国信息通信研究院/中国泰尔实验室	北京市海淀区花园北路 52 号
3.	170009010819	国家电话机质量监督检验中心	北京市海淀区花园北路 52 号
4.	170009010911	信息产业邮电工业产品质量监督检验中心	北京市海淀区花园北路 52 号
5.	170009011068	信息产业图文通信设备质量监督检验中心	北京市海淀区花园北路 52 号
6.	170009011289	信息产业北京移动通信设备质量监督检验中心	北京市海淀区花园北路 52 号
7.	170009011635	信息产业北京电话交换设备质量监督检验中心	北京市海淀区花园北路 52 号
8.	170009012175	信息产业通信电磁兼容质量监督检验中心	北京市海淀区花园北路 52 号
9.	170009012615	信息产业通信软件测评中心	北京市海淀区花园北路 52 号
10.	170009013675	国家物联网通信产品质量监督检验中心	北京市海淀区花园北路 52 号
11.	170009012246	中国电子产品可靠性与环境试验研究所（（工业和信息化部电子第五研究所）（中国赛宝实验室））	广东省广州市增城区朱村街朱村大道西 78 号
12.	160021020992	公安部安全与警用电子产品质量检测中心	北京市海淀区首都体育馆南路 1 号
13.	160021022236	公安部特种警用装备质量监督检验中心	北京市海淀区首都体育馆南路 1 号
14.	160021022463	国家安全防范报警系统产品质量监督检验中心（北京）	北京市海淀区首都体育馆南路 1 号
15.	210020024472	公安部第一研究所	北京市市辖区海淀区首都体育馆南路 1 号
16.	170009020967	公安部安全防范报警系统产品质量监督检验测试中心	上海市徐汇区岳阳路 76 号
17.	170021022464	国家安全防范报警系统产品质量监督检验中心(上海)	上海市徐汇区岳阳路 76 号
18.	170009021800	公安部计算机信息系统安全产品质量监督检验中心	上海市徐汇区岳阳路 76 号

19.	170009022408	**公安部信息安全产品检测中心**	上海市徐汇区岳阳路 76 号
20.	170009023583	**国家网络与信息安全产品质量监督检验中心**	上海市徐汇区岳阳路 76 号
21.	160008220102	**国家日用电器质量监督检验中心**	广东省广州市黄埔区科学城开泰大道天泰一路 3 号
22.	160008222171	**威凯检测技术有限公司**	广东省广州市黄埔区科学城开泰大道天泰一路 3 号
23.	210009344471	**博鼎实华（北京）技术有限公司**	北京市海淀区建材城中路 27 号 N5 楼
24.	160009113420	**国家信息安全产品质量监督检验中心（北京）**	北京市朝阳区朝外大街甲 10 号中认大厦
25.	170009010178	北京泰瑞特检测技术服务有限责任公司/国家广播电视产品质量监督检验中心	北京市朝阳区酒仙桥北路乙 7 号
26.	170009012149	信息产业广播电视产品质量监督检验中心	北京市朝阳区酒仙桥北路乙 7 号
27.	180009112775	国家数字电子产品质量监督检验中心	广东省深圳市南山区西丽街道同发路 4 号
28.	210009344449	工业互联网创新中心（上海）有限公司	上海市浦东新区金港路 766 号四号楼
29.	160009012247	信息产业信息安全测评中心	北京市海淀区北四环中路 211 号
30.	170009012416	国家应用软件产品质量监督检验中心	北京市市辖区海淀区东北旺西路 8 号中关村软件园 3A 楼
31.	170010110145	中家院（北京）检测认证有限公司（中国家用电器检测所）/国家家用电器质量监督检验中心	北京市大兴区经济技术开发区博兴八路 3 号
32.	170010263004	国家轻工业家用电器质量监督检测中心	北京市大兴区经济技术开发区博兴八路 3 号
33.	170010264040	国家智能家居质量监督检验中心	北京市大兴区经济技术开发区博兴八路 3 号
34.	180008221885	上海电器设备检测所有限公司	上海市普陀区武宁路 505 号
35.	180008223354	国家智能电网用户端产品（系统）质量监督检验中心	上海市普陀区武宁路 505 号
36.	180009252345	电力工业通信设备质量检验测试中心	北京市海淀区清河小营东路 15 号
37.	180020252130	国网电力科学研究院有限公司实验验证中心	江苏省南京市江宁区诚信大道 19 号
38.	180020252265	中国电力科学研究院有限公司	北京市海淀区清河小营东路 15 号
39.	190008224237	国家机器人质量监督检验中心	上海市普陀区武宁路 505 号
40.	190008224238	国家汽车电气化产品及系统质量监督检验中心	上海市普陀区武宁路 505 号
41.	190020254257	国家输配电安全控制设备质量监督检验中心	江苏省南京市江宁区诚信大道 19 号
42.	200009014368	国家工业互联网系统与产品质量监督检验中心	上海市市辖区普陀区上海市普陀区武宁路 505 号

43.	200021014043	国家工业控制系统与产品安全质量监督检验中心	北京市石景山区鲁谷路 35 号
44	180009011125	信息产业数据通信产品质量监督检验中心	北京市海淀区学院路 40 号

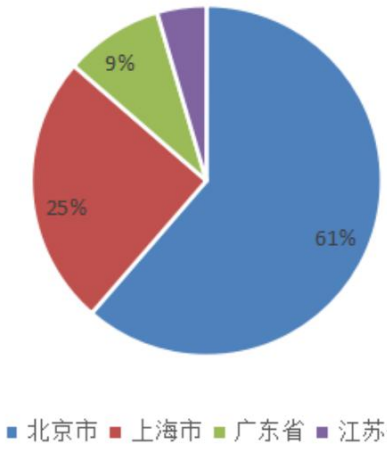

图 8 具有 CMA 资质的人工智能安全及相关检验检测机构地域分布

2. CNAS 认可的实验室

我国现有获得 CNAS 认可的实验室 10645 家，其中属于企业内部实验室约 1830 家。据不完全统计，在人工智能安全领域具有检测能力的实验室有 42 家，公安体系 3 家，工信体系 4 家，如表 3 所示。从地域分布上看，主要集中在北京、上海、广东、江苏和浙江，如图 9 所示。

表 3 CNAS 认可人工智能安全及相关实验室

序号	注册编号	机构名称	地域	其他名称
1.	L10453	国家工业信息安全发展研究中心（工业和信息化部电子第一研究所）检查评估所	北京市	
2.	L0462	中国赛宝实验室/工业和信息化部电子第五研究所/中国电子产品可靠性与环境试验研究所	广东省广州市、浙江省宁波市	国家环境试验设备质量监督检验中心，国家通用电子元器件及产品质量监督检验中心，国家集成电路产品质量监督检验中心，国家嵌入式软件产品质量监督检验中心，国家卫星导航及应用产品质量监督检验中心，广东省质量监督软件产品检验站国家环境综合试验设备计量站，国家汽车电子产品质量监督检

				验中心，国家无人机系统质量监督检验中心，广东省质量监督智能家用电子电器产品检验站（广州）
3.	L0357	**中国电子技术标准化研究院赛西实验室**	北京市	国家数字音视频及多媒体产品质量监督检验中心 信息处理产品标准符合性检测中心 电子工业安全与电磁兼容检测中心 工业和信息化部电子计量中心 工业和信息化部数字电视标准符合性检测中心 国家大数据系统产品质量监督检验中心 国防科技工业微电子元器件一级计量站 国家绿色电池质量监督检验中心 国家虚拟现实/增强现实产品质量监督检验中心
4.	L0570	**中国信息通信研究院泰尔实验室**	北京市、河北省保定市	中国泰尔实验室国家电话机质量监督检验中心，信息产业北京移动通信设备质量监督检验中心，信息产业图文通信设备质量监督检验中心，信息产业北京电话交换设备质量监督检验中心，信息产业通信电磁兼容质量监督检验中心，信息产业通信软件测评中心，信息产业邮电工业产品质量监督检验中心，信息产业通信设备抗震性能质量监督检验中心，国家物联网通信产品质量监督检验中心
5.	L2110	**公安部交通管理科学研究所交通安全产品质量监督检测中心**	江苏省无锡市	国家道路交通安全产品质量监督检验中心，公安部交通安全产品质量监督检测中心
6.	L0653	**公安部第三研究所安全防范与信息安全产品及系统检验实验室**	上海市、北京市	国家安全防范报警系统产品质量监督检验中心（上海），公安部安全防范报警系统产品质量监督检验测试中心，公安部计算机信息系统安全产品质量监督检验中心，公安部信息安全产品检测中心，公安部信息安全等级保护评估中心国家网络与信息系统安全产品质量监督检验中心上海市安全防范产品质量监督检验站
7.	L0531	**公安部第一研究所安全与警用电子产品质量检测中心**	北京市、浙江省杭州市	国家安全防范报警系统产品质量监督检验中心（北京），公安部特种警用装备质量监督检验中心，公安部安全与警用电子产品质量检测中心
8.	L6398	**中国网络安全审查技术与认证中心**	北京市	国家信息安全产品质量监督检验中心（北京）
9.	L0579	深圳市计量质量检测研究院	广东省深圳市	国家高新技术计量站国家数字电子产品质量监督检验中心，国家质量监督检验检疫总局深圳计量检定站国家体育用品质量监督检验中心（广东），国家营养食品质量监督检验中心（广东），国家分布式光伏发电系统质量监督检验中心（广东），国家环保产品质量监督检验中心（广东），中国轻工业联合会

				家具质量监督检测深圳站广东省质量监督眼镜检验站（深圳），广东省质量监督自行车检验站，广东省质量监督综合布线系统检验站，广东省质量监督电磁兼容检验站，广东省质量监督皮革制品检验站，广东省质量监督生态纺织服装产品检验站（深圳），广东省质量监督钟表检验站（深圳），广东省质量监督食品检验站（深圳），广东省质量监督家具检验站（深圳），广东省质量监督环保节能产品（安全性能）检验站（深圳），广东省质量监督学生用品检验站（深圳），广东省质量监督名贵饰品检验站（深圳）
10	L15379	山东省标准化研究院条码印刷品质量检验站	山东省济南市	
11	L13377	国网汇通金财（北京）信息科技有限公司信息通信测试与仿真中心	北京市西城区	
12	L1379	国网电力科学研究院有限公司实验验证中心	江苏省南京市	电力工业电力系统自动化设备质量检验测试中心，国家电网公司自动化设备电磁兼容实验室，电力工业通信设备质量检验测试中心，国家输配电安全控制设备质量监督检验中心
13	L0699	中国电力科学研究院有限公司	北京市、河北省霸州市、湖北省武汉市、江苏省南京市	国家风电技术与检测研究中心，电力工业电力设备及仪表质量检验测试中心，电力工业电力系统自动化设备质量检验测试中心，电力工业电力工程材料部件质量检验测试中心，电力工业电力及通信混泥土电杆质量检验测试中心，电力工业通信设备质量检验测试中心，电力工业电气设备质量检验测试中心
14	L0128	上海市质量监督检验技术研究院	上海市	国家电光源质量监督检验中心(上海)，国家灯具质量监督检验中心，国家食品质量监督检验中心(上海)，国家保洁产品质量监督检验中心，国家家具质量监督检验中心，国家建筑材料及装饰装修材料质量监督检验中心，国家电器能效与安全质量监督检验中心，国家轻工业建筑五金质量监督检测中心，国家轻工业灯具质量监督检测中心，国家轻工业家具质量监督检测中心，中国商业联合会交电家电商品质量监督检验测试中心(上海)，化学工业鞋类质量监督检验中心，国家日用消费品质量监督检验中心，国家智能电网分布式电源装备质量监督检验中心（上海）

15	L0768	北京信息安全测评中心	北京市	
16	L10857	国家信息中心电子政务信息安全等级保护测评中心	北京市	
17	L1145	上海电器设备检测所有限公司	上海市、浙江省湖州市	国家低压电器质量监督检验中心，国家中小电机质量监督检验中心，国家智能电网用户端产品（系统）质量监督检验中心，国家机器人质量监督检验中心，国家汽车电气化产品及系统质量监督检验中心，国家工业互联网系统与产品质量监督检验中心，机械工业船用电气设备试验站机械工业产品电磁兼容性质量监督检测中心
18	L0943	北京软件产品质量检测检验中心	北京市	国家应用软件产品质量监督检验中心
19	L0793	中家院（北京）检测认证有限公司（中国家用电器检测所）	北京市、浙江省慈溪市	国家家用电器质量监督检验中心，国家智能家居质量监督检验中心
20	L14624	博鼎实华（北京）技术有限公司	北京市	
21	L0500	北京泰瑞特检测技术服务有限责任公司	北京市	国家广播电视产品质量监督检验中心
22	L0293	华北计算技术研究所计算机测评中心	北京市	信息产业信息安全测评中心
23	L8484	安徽继远检验检测技术有限公司	安徽省合肥市	
24	L1066	北京通和实益电信科学技术研究所有限公司威尔克通信实验室	北京市	信息产业数据通信产品质量监督检验中心
25	L14481	工业互联网创新中心（上海）有限公司	上海市	
26	L12381	国信金宏信息咨询有限责任公司	西藏自治区拉萨市	
27	L9000	北京亮点软测科技有限公司	北京市	
28	L0160	南京市产品质量监督检验院	江苏省南京市	国家建材产品质量监督检验中心（南京），国家金银制品质量监督检验中心（南京），国家加工食品及食品添加剂质量监督检验中心（南京），江苏省智能电网应用产品质量监督检验中心 国家软件产品质量监督检验中心（江苏）
29	L0768	北京信息安全测评中心	北京市	
30	L10857	国家信息中心电子政务信息安全等级保护测评中心	北京市	
31	L0754	上海市信息安全测评认证中心	上海市	

32	L0943	北京软件产品质量检测检验中心	北京市	国家应用软件产品质量监督检验中心
33	L5302	国家信息技术安全研究中心信息安全特种技术检测实验室	北京市	
34	L0095	威凯检测技术有限公司	广东省广州市	国家日用电器质量监督检验中心，机械工业日用电器产品质量监督检测中心，机械工业分马力电机产品质量监督检测中心，机械工业电工产品环境适应性检测中心，机械工业日用电器电磁兼容性监督检测中心，广东省质量监督分马力电机检验站
35	L0116	浙江方圆检测集团股份有限公司	浙江省杭州市	国家化学建材质量监督检验中心 国家皮革质量监督检验中心（浙江） 国家电器安全质量监督检验中心（浙江） 国家预包装食品质量监督检验中心（浙江） 国家电子商务消费品质量监督检验中心（浙江） 浙江省黄金珠宝饰品质量检验中心 浙江省电动车辆产品质量检验中心 浙江省智能技术质量检验中心 浙江省安全技术质量检验中心 浙江省低压电器产品质量检验中心
36	L0293	华北计算技术研究所计算机测评中心	北京市	信息产业信息安全测评中心
37	L9160	山东智慧云测信息科技有限公司	山东省济南市	
38	L5226	成都泰瑞通信设备检测有限公司	四川省成都市	信息产业有线通信产品质量监督检验中心 国家传送网产品与系统安全质量监督检验中心
39	L14922	杭州中正检测技术有限公司	浙江省杭州市	
40	L13377	国网汇通金财（北京）信息科技有限公司信息通信测试与仿真中心	北京市	
41	L13367	山东卓朗检测股份有限公司	山东省济宁市	
42	L1379	国网电力科学研究院有限公司实验验证中心	江苏省南京市	电力工业电力系统自动化设备质量检验测试中心，国家电网公司自动化设备电磁兼容实验室，电力工业通信设备质量检验测试中心 国家输配电安全控制设备质量监督检验中心

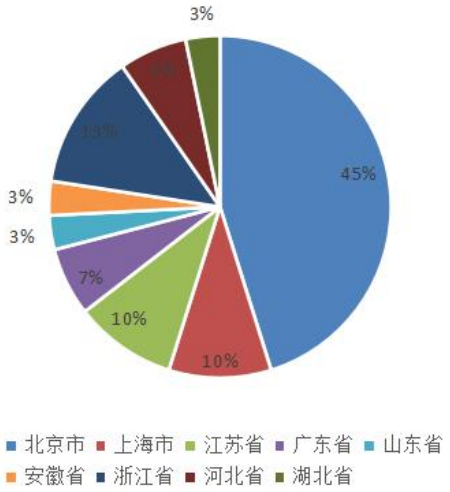

图 9 具有 CNAS 资质的人工智能安全及相关实验室地域分布

3. CNAS 认可的认证机构

我国现有能够开展认证活动的机构 888 家，开展强制性产品认证的机构 116 家，开展强制性认证的指定实验室 740 家。据不完全统计，具有人工智能产品认证能力的机构有 10 家，工信体系 3 家，公安体系 1 家，金融体系 2 家，如表 4 所示。从依据标准的类别看，主要为技术规范、地方标准、团体标准，人工智能相关产品国标认证明显不足。从认证产品上看，主要为人脸识别系统、智能音箱、智能手机和智能门锁产品等。中互金认证有限公司、中国安全技术防范认证中心、深圳市计量质量检测研究院在数据安全方向率先开启认证。中国网络安全审查技术与认证中心、北京赛西认证有限责任公司在可信环境和人脸识别安全具有标杆效应。

表 4 CNAS 认可人工智能安全及相关认证机构

序号	认证机构	行政区划	依据标准	依据标准名称	依据标准编号
1.	北京赛西认证有限责任公司	北京市东城区	技术规范	《人脸识别系统认证技术规范》	CTS 039-2021
2.	广州赛宝认证中心服务有限公司	广东省广州市增城区	技术规范	《智能音箱语音交互性能优选认证技术规范》	CTS CEPREI-0
3.	泰尔认证中心有限公司	北京市西城区	团体标准	《智能门锁信息安全技术要求和评估方法》	OT_GR T/TAF 030-2019
4.	公安部第一研究所（中国安全技术防范认证中心）	北京市海淀区	技术规范	《汽车行驶记录仪用防护存储器数据安全性认证检测技术规范》	CTS CSP-J005
5.	中国网络安全审查技术与认证中心	北京市朝阳区	行业标准	《移动终端支付可信环境技术规范》	JR/T 0156-2017
6.	中互金认证有限公司	天津市滨海新区	技术规范	《联邦学习数据安全技术规范》	CTS Q/NIFC 001-2021
7.	北京国家金融科技认证中心有限公司	北京市西城区	行业标准	《金融科技创新安全通用规范》《金融科技创新风险监控规范》	JR/T 0199
8.	深圳市计量质量检测研究院	广东省深圳市南山区	技术规范	《信息安全技术 移动智能终端应用软件安全技术要求和测试评价方法》、《信息安全技术 个人信息安全规范》	GB/T 34975 、GB/T 35273
9.	杭州万泰认证有限公司	浙江省杭州市滨江区	地方标准	《深圳标准先进性评价细则-Android 智能手机》	SZJG SSAE-A12-001：2018
10.	深圳华测国际认证有限公司	广东省深圳市宝安区	地方标准	《深圳标准先进性评价细则-Android 智能手机》	SZJG SSAE-A12-001：2018

（四）安全检测工具现状

近年来，美国发布《保持美国在人工智能领域的领导地位》、《国家人工智能研发战略规划》、《美国在人工智能领域的领导地位：联邦政府参与开发技术标准和相关工具的计划》等多项文件，表明美国高度重视人工智能治理相关的标准和工具研发，把安全检测工具作为科技竞争

的重要方向。我国发布《新一代人工智能发展规划》、《促进新一代人工智能产业发展三年行动计划（2018-2020年）》等多项文件，支持研发人工智能算法与平台安全性测试评估的方法、技术、规范和工具集。由于人工智能安全测评处于理论研究和工具研发阶段，国内外针对人工智能安全测评的工具较少，国外主要集中在对算法模型的透明性、公平性、可解释性评估，典型的检测工具有谷歌ModelCards、IBM Adversarial Robustness Toolbox、微软Counterfit，国内主要集中在算法模型安全、对抗攻击的评估、学习框架安全性评估，典型的检测工具有百度advbox、华为MindArmour、瑞莱RealSafe、360AIFater，如表5所示。

表5典型人工智能安全检测工具

序号	企业	关键技术	检测工具	实践应用
1	谷歌	透明性AI技术 公平性AI技术	ModelCards FairnessIndicators	COVID-19、 名人识别API、 谷歌翻译
2	IBM	对抗攻击AI技术 可解释性AI技术 公平性AI技术 模型推断检测技术 数据投毒检测	IBMAdversarialRobustnessToolbox AIExplainability360 AIFairness360	KMPG毕马威、 RegionsBank
3	微软	Fairlearn和AI公平性检查表 InterpretML和数据集的数据表	Counterfit Fairlearn InterpretML SmartNoise ErrorAnalysis	对话机器人
4	百度	对抗攻击技术 黑白盒攻击技术	advbox	机器学习模型
5	华为	鲁棒性增强技术 对抗攻击技术	MindArmour	机器学习模型
6	瑞莱	黑盒查询攻击技术 黑盒迁移攻击技术 漏洞检测技术	realsafe	机器学习模型

		模型安全性测评 模型鲁棒性提升		
7	360	云端机器学习框架 安全性评估技术 终端机器学习框架 安全性评估技术	AIFater	机器学习框架

1. 谷歌 Model Cards、Fairness Indicators

ModelCards 和 Fairness Indicators 是谷歌研发的观察机器学习模型透明度信息模型卡任务工具和用于计算和可视化分类模型中常见的公平性指标的工具。ModelCards 提供模型来源、使用方法和道德规范相关评估信息的结构性框架，并且还会提供模型使用限制建议。Fairness Indicators 可根据已部署生产的机器学习(ML)检测工具模型的上下文来检测和修正该模型中的偏差。

2. IBM Adversarial Robustness Toolbox 、 AI Explainability 360、AI Fairness 360

Adversarial Robustness Toolbox 是 IBM 研究团队开源的用于检测模型及对抗攻击的工具箱，为开发人员加强 AI 模型被误导的防御性，目前支持 TensorFlow 和 Keras 框架。AI Explainability360（AIX360）是模型可解释性工具，涵盖了八种前沿的可解释性方法和两个维度评价矩阵。同时还提供了有效的分类方法引导各类用户寻找最合适的方法进行可解释性分析。AI Fairness360（AIF360）是模型公平性工具包，可帮助检测和减轻整个 AI 应用程序

生命周期中机器学习模型的偏差。用于数据集和模型的全面指标集，以测试偏差。

3. 微软 Counterfit、Fairlearn、InterpretML、SmartNoise、Error Analysis

Counterfit 微软开源的 AI 检测工具安全风险评估工具，可以评估机器学习系统的安全性，确保其业务中使用的算法是可靠和可信赖的。InterpretML、Fairlearn、SmartNoise 和 ErrorAnalysis 是微软推出的四个响应式 AI 工具包，可以用于机器学习模型预测、评估机器学习算法公平性、保护敏感数据、识别和诊断模型的不精确性。

4. 百度 advbox

AdvBox 是百度安全实验室研发的 AI 模型安全工具箱，目前原生支持主流机器学习平台，同时支持 GraphPipe，屏蔽了底层使用的深度学习平台，用户可以零编码，实现对抗样本生产，模型的对抗样本测试以及模型的黑白盒攻击等功能。通过提供对抗样本生成工具帮助用户评估深度学习模型，工具包提供了一些评价指标，例如攻击成功率等，使用者能直观地查看模型在某种攻击下的防御成功率。

5. 华为 MindArmour

MindArmour 是华为研发的 AI 模型安全测试工具，主要针对对抗性样本引起的攻击行为，包括了对抗样本生成、对抗样本检测、模型防御、模型评估四个子模块。对抗样本生成模块支持安全工程师快速高效生成对抗样本，用于攻击 AI 模型；对抗样本检测、防御模块支持用户检测过滤对抗样本、增强 AI 模型对于对抗样本的鲁棒性；评估模块提供多种指标全面评估对抗样本攻防性能。MindArmour 为 AI 模型安全研究和 AI 应用安全提供重要支撑。

6. 瑞莱 RealSafe

RealSafe 是瑞莱研发的针对 AI 在极端和对抗环境下的算法安全性检测与加固的工具。RealSafe 主要支持两大功能模块：模型安全测评和防御解决方案。模型安全测评主要为用户提供 AI 模型安全性测评服务。目前已支持黑盒查询攻击方法与黑盒迁移攻击方法。防御解决方案则是为用户提供模型安全性升级服务，目前 RealSafe 平台支持五种去除对抗噪声的通用防御方法，可实现对输入数据的自动去噪处理，破坏攻击者恶意添加的对抗噪声。随着模型攻击手段在不断复杂扩张的情况下，RealSafe 平台还持续提供广泛且深入的 AI 防御手段，帮助用户获得实时且自动化

的漏洞检测和修复能力。

7. 360 AIFater

AIFater 是 360 研发的针对机器学习框架的安全性评测工具，主要具备两大功能：云端框架安全性评测和终端框架安全性评测。AIFater 基于丰富的专家知识库，结合静态分析与动态分析的优势，对机器学习框架进行全面、系统、自动化的安全性测试，具备对机器学习框架持续性的攻击面测绘、风险发现、漏洞分析的能力。AIFater 不仅保障了开发者在使用机器学习框架时的开发环境安全，同时保障了用户在使用 AI 服务时的隐私安全。

四、人工智能安全测评建议

我国人工智能发展正进入技术创新迭代持续加速和融合应用拓展深化的新阶段，对各类人工智能系统的安全测评充分反映出当前人工智能技术的应用仍存在非常大的风险。当前阶段，妥善评估并解决由人工智能安全问题引发的相关威胁已经刻不容缓。需加深国际治理探讨交流，制定适宜国情监管政策，加快关键核心技术研发，统筹完善安全标准体系，加强智能检测工具研发，建设精准监测预警平台，加速检测认证能力建设，审核一批权威测评机构。

（一） 加深国际治理探讨交流，制定适宜国情监管政策

人工智能的发展需要全球化，人工智能监管也离不开国际交流。我国现有的监管主要集中在数据安全和网络安全领域，其他领域还没有形成系统健全的监管框架，不利于营造透明有序的创新环境。我国提出技术创新与安全发展并重的人工智能发展理念，积极参与国际人工智能治理相关工作，提升我国国际规则制定的参与度和影响力。结合我国人工智能产业的发展特点，借鉴欧盟的监管思路，探索我国人工智能安全分类分级监管机制，明确人工智能系统禁止和限制使用规则，构建人工智能监管组织架构，明确权责分配和设置严厉的处罚措施，建立人工智能系统备案制度，有序推进人工智能系统备案工作，高风险人工智能系统需在监管部门注册备案，便于接受监管，遵循风险分级、最小干预原则，实现以安全促发展的目标。

持续推进监管模式创新，通过“自监管+外部监管”的模式，构建多元主体协同合作的人工智能监管体系。强化企业主体责任和责任意识，鼓励企业建立人工智能安全责任制度和科技伦理审查制度，加强风险防控和隐患排查治理，提升企业应对人工智能安全突发事件的能力和水平。倡导企业先通过企业伦理规则和内部审查机制进行自监管，再通过外部监管部门依据法律法规抽检，引导企业逐

步实现先安全评估再上线运行。鼓励社会监督参与，鼓励广大人民积极参与人工智能安全监督治理工作，切实加强政府、企业、行业组织和广大人民群众间的信息交流和有效沟通，政府积极受理人民群众建议与投诉，企业自觉接受社会监督，及时做好结果反馈。

（二） 加快关键核心技术研发，统筹完善安全标准体系

技术研发和标准研制需同步推进，技术研发保证技术内生的安全可靠，标准研制规范技术应用的合规可控，二者相辅相成，密不可分。当前人工智能技术的发展一共经历了两代。为解决第一代和第二代人工智能存在的不可解释、鲁棒性差、不安全、不可信、不可靠、不可扩展等缺点，提出第三代人工智能技术研究。除此之外，加强人工技能基础技术研究，包括小样本学习、迁移学习、联邦学习等，加快人工智能共性技术研究，包括高端智能芯片、深度学习框架、核心算法等，加速人工智能应用技术研究，包括多模态感知技术、人工智能与区块链融合技术、人工智能与量子信息交叉技术等。

需要统筹规划人工智能安全标准体系，鼓励标准化组织、行业协会、产业联盟等积极参与国际标准化工作，尤其在基础安全、数据、模型算法、计算设施安全等方面的标准研究。研制人工智能通用安全参考架构，保障人工智

能系统安全所涉及的利益相关方、相关方角色活动、安全功能视图等，为后续具体规范人工智能各方面安全提供参考。研制人工智能开源框架、人工智能服务器、人工智能开发平台、AI 计算平台方面安全标准，为人工智能核心资产提供安全的运算环境以及针对性的安全保护措施。研制 AI 模型的安全保护实施指南标准，在机器学习算法安全评估规范的基础上，补充 AI 模型的内生因素、运行环境因素等各方面内容以完善前面的 AI 模型安全保护指南。研制机器学习训练和测试数据集安全性评估指南，以及人工智能信息系统中的数据安全治理指南，对既有标准法规提出的数据集均衡无偏见，欧盟 AI 法规提出的数据安全治理等内容给出实施指南方面标准补充。除此之外，还需深化人工智能应用安全标准研究，针对已有标准完善智能安全要求，优先针对标准化需求迫切且应用较成熟，或应用较敏感的领域进行标准研究。

（三） 加强智能检测工具研发，建设精准监测预警平台

密切跟踪人工智能安全攻防、安全检测等关键技术进展，及时掌握攻击和检测原理，积极开展人工智能安全评估，建立人工智能测试评估体系，研发智能产品和服务的可靠性、安全性等检测工具，推进人工智能安全测评平台落地推广，对重点高风险智能技术、产品和服务的功能、

性能、安全性等进行评估，提升人工智能技术、产品和服务质量。面向智能语音、图像识别、智能机器人、自动驾驶、工业质检领域建设第三方试点测试平台并开展评估测评服务。

面向人工智能产业发展，整合监测资源，针对重点行业建立人工智能系统安全监测预警平台，理解并解决人工智能的伦理法律和社会影响、确保人工智能系统安全且可靠、开发可共享的用于人工智能系统训练和测试的公共数据环境、建立衡量评估人工智能技术的基准和标准、更好地了解人工智能研发人员需求、扩大人工智能领域的公私伙伴关系等，并对实施进展情况开展阶段性评估。基于人工智能安全测评，完善应急管理体系和能力，防范化解人工智能重大安全风险、及时应对处置各类相关灾害事故，防范算法滥用风险，维护网络空间正常秩序，防范利用人工智能算法模型干扰社会舆情民意、打压行业竞争对手、损害公民合法权益等行为，防范人工智能算法模型滥用带来的意识形态、经济发展和社会管理等方面的风险隐患。

（四） 加速检测认证能力建设，审核一批权威测评机构

构建与国际接轨的我国人工智能检测认证体系，对高风险人工智能应用在投放市场前开展第三方符合性评估活动，评估安全风险等级。加强产品质量测试、安全评估、

检验认证，建设安全标准计量、认证认可、检验检测、试验验证等产业技术示范应用平台，建立完善可对智能设备设计、开发、测试、运营全生命周期进行安全测评能力，培育发展一批专业检测认证机构。并依据监管要求对必要的关键基础设施设备及相关重要设备开展安全检测，联合企业共同制定解决方案，并督促实施。

完善第三方测评机构审核流程，加强第三方测评机构准入机制，对不符合资质要求的面向社会开展第三方的测评机构进行处罚和取缔，提升检验检测认证行业的检验检测认证水平。测评机构对经过一定程序开展评估并符合要求的人工智能系统可颁发证书，规定证书的有效期限，实现统一符合性评定方式对人工智能产品或服务的监管，加强人工智能发展与监管领域的国际交流，提升国际相关标准的话语权，为我国企业提供更便利、更就近的合规评估服务，尽可能降低企业进入海外市场的门槛。

附件 1 技术安全防护实践案例

（一） 瑞莱“深度伪造”安全实践

“深度伪造”是指利用深度学习的技术对图片、视频等内容进行面部替换、表情编辑等篡改行为。自 2017 年出现以来，“深度伪造”伴随着人工智能技术在音视频领域的快速发展及应用普及，以 DeepFakes、GAN 以及各种移动端 APP 工具为代表的深伪技术日新月异，“深度伪造”生成内容的门槛越来越低，且生成的内容逼真，具有伪装性强、欺骗性高、危害性大的特点。

1. 深度伪造内容检测方案 DeepReal

基于深度伪造技术带来的风险与危害，瑞莱智慧推出深度伪造内容的检测方案 DeepReal。针对视频、图像、音频的多模态信息，利用深度学习的方法，实现对内容真实或伪造的判断，支持输出可视化的分析结果与多维度的检测报告。DeepReal 深度伪造检测的方法如下：

基于伪造特征提取的检测方法：通过分析空域与频域特征、音视频多模态信息的统一表征、动态生物信号等关键信息，实现对深度伪造内容的检测。

基于神经网络设计的检测方法：在深度伪造神经网络模型中加入先验知识，通过设计高效的神经网络结构，提

升检测的鲁棒性和高效性。

基于模型训练优化的检测方法：基于机器学习泛化理论、元学习等方法，实现在开放域、小样本场景下的模型迭代，提升面对新型数据的泛化能力。

多检测模型集成：集成多种深度伪造检测模型，基于模型融合的方法和策略，整合不同检测模型的优势，保障整体方案的可靠性。

基于上述技术，可以实现对面部替换、表情操纵、全脸合成、表情编辑等多种深度伪造方法的甄别，应用于检测网络内容合规性、提升人脸验证安全性、检测物证材料真实性、打击侵权诈骗行为等业务场景中。

2. DeepReal 应用：业务数据的真实性检测

瑞莱智慧推出的深度伪造内容检测方案支撑某监管单位对业务数据的真实性检测。在产业实践中发现，网络中传播的深度伪造的内容采用的合成方法丰富，且在传播过程中音视频文件会发生不同的质量变化与扰动，这些因素均影响深度伪造检测方法的实战效果。

面对实网中复杂的对抗数据环境，采用对抗学习等方法提升检测方法的鲁棒性水平；面对新型方法的伪造数据，采用基于小数据量的快速迭代技术提升检测方法的泛化性水平；面向实际应用中的流式数据，采用离线与在线

学习相结合的方式，提升检测方法的响应能力。基于以上全面的技术方案，提供应对实网环境对抗变化的工业级检测能力。

在深度伪造检测基础上，用户仍有进一步追溯造假责任的需求。针对这一问题，也支撑开展深度伪造音视频的溯源、伪造时空推演等方面的研究，为用户提供更多的可解释性线索。

（二） 华为“对抗样本”安全实践

“对抗样本”是指通过给样本添加一些轻微的扰动，会导致基于神经网络的人工智能模型分类错误。将以深度学习为代表的机器学习模型应用到工程实践时，会因以上原因产生安全问题，尤其在对安全要求较高的工业应用中，如金融、自动驾驶等领域，恶意的对抗样本攻击会造成严重的人身财产损失。将深度学习模型用于工业实践，进行模型对对抗样本的鲁棒性增强与测试，尽可能保证模型在使用过程中的安全。

1. 对抗样本测试工具 MindArmour

对抗样本对神经网络能形成有效攻击的原因之一是神经网络学习的输入与输出映射在很大程度上是相当不连续的，我们可以通过应用某些难以感知的扰动来使网络输出

错误的预测，而扰动是通过最大化网络的预测误差来发现。华为基于自研的开源深度学习框架 MindSpore 以及业务实践，形成了一套完整的模型鲁棒性增强与测试方案——MindArmour。MindArmour 提供：涵盖黑白盒对抗攻击、成员/属性推理攻击、数据漂移等测试数据的生成方法；提供基于覆盖率的 Fuzzing 测试流程，可灵活制定测试策略与指标；提供对抗训练、输入重建在内的常见对抗样本检测和模型鲁棒性增强方法。具体的解决方案框架，如图 10 所示，包括：用户交互模块、对抗样本生成模块、对抗样本检测模块、对抗样本防御模块以及测试评估模块，用户可通过输入模块自定义测试的种子数据集、被测的模型、训练参数配置与防御参数配置，进行对抗样本测试或模型鲁棒性增强，进行评估后通过输出接口获取测试结果。

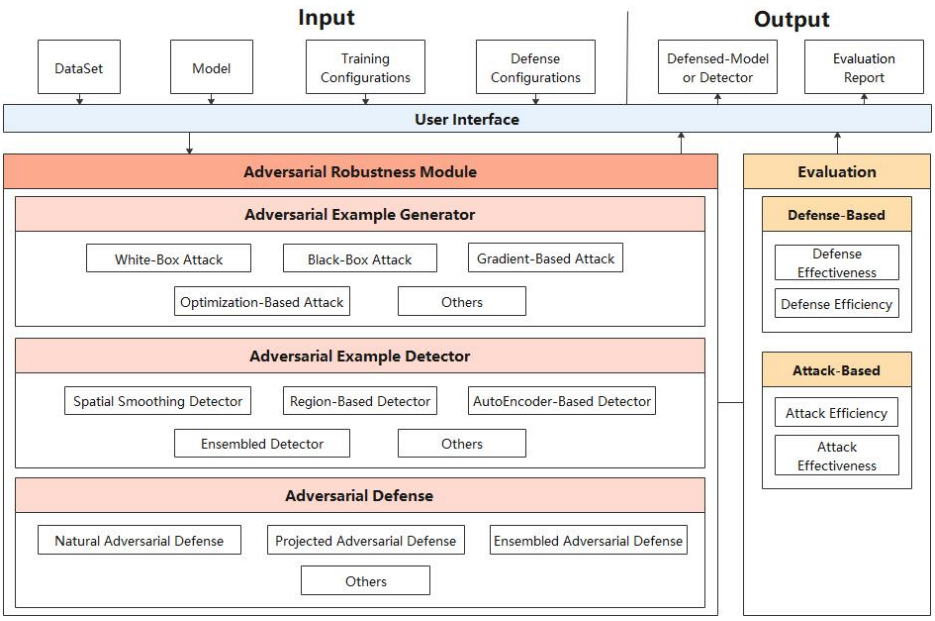

图 10 MindArmour 对抗鲁棒性解决方案

2. MindArmour 应用：OCR 服务端到端对抗测试

OCR 是从图像中识别和提取文字的技术，包含了多个 AI 模型，可以将图像中的文字直接转换为计算可直接处理的数据文件，应用场景广泛，包括金融票据、物流行业等，也可为下游任务提供文字数据，机器翻译。

光学字符识别（OCR）服务的处理过程包含了多个 AI 模型，系统主要的威胁来源于对抗输入扰动（即对抗攻击）和自然扰动两个方面。减轻这些威胁的措施需要提高 OCR 服务中 AI 模型的鲁棒性，并且可以通过测试的手段进行定量的评估。

对于恶意输入扰动，可以建立如下的威胁模型：一是攻击者的目标，聚焦于在推理阶段，利用对抗攻击的方式去影响 AI 服务的预测结果，而具体的做法就是操作输入图像，如增加扰动。二是攻击者的能力，攻击者是否可以直接操纵 OCR 系统的输入数据，或者攻击者对输入数据增加的扰动是否容易被感知，如增加扰动后，是否可以被人眼识别出来，可以用 L2 范数作为扰动大小的指标。三是攻击者的知识背景，通过攻击对模型的了解程度可以将其分为知道模型结构与参数的白盒攻击和对模型结构参数不清楚的黑盒攻击。在黑盒攻击中，攻击者也无法知道 OCR 服务内部处理的流程与细节，因此只能进行端到端的攻击。

测试包含对 OCR 服务端到端的测试和对 OCR 服务中的模型进行模块测试，测试每个模型的脆弱性。对于自然扰动的产生，不取决于特定的 AI 模型，可以用单个模块的黑盒测试中，也可以用在端到端的测试中。对于对抗扰动，在模型的开发阶段，开发者需要对模型进行白盒测试，并在端到端的测试中使用黑盒方法产生测试样本。OCR 测试模型包含角度检测模型、文本检测模型、反转检测模型和文字识别模型，如图 11 所示的测试方案，一个 OCR 服务安全测试，分为模型测试和端到端的测试。模型测试需要测试中所述的四个模型，测试方法要有自然扰动和对抗扰动；端到端的测试以自然扰动和黑盒对抗方法为主。

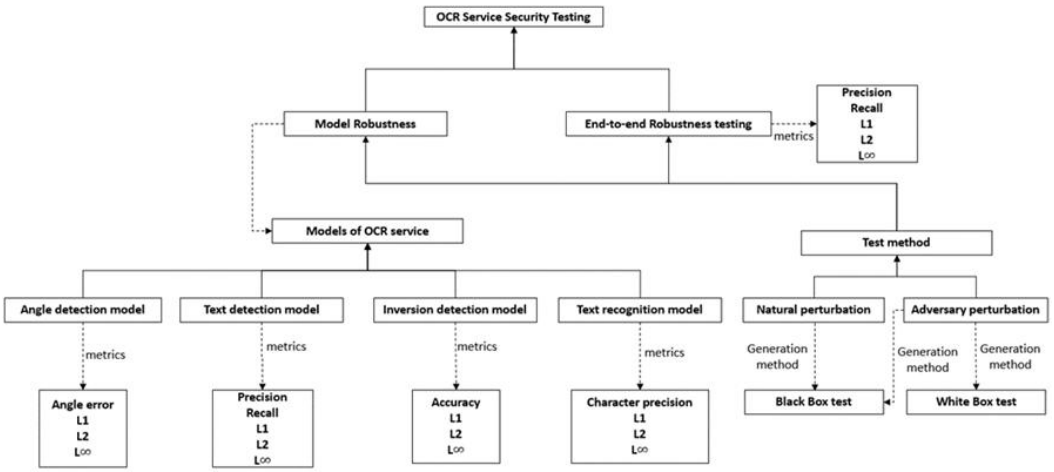

图 11 OCR 测试方案

（三） 百度“数据劫持”安全实践

“数据劫持”是指在数据采集、传输、使用过程中，通过对设备硬件、操作系统、应用程序或网络链路，使用

注入、替换等技术拦截或篡改数据。例如在人脸识别场景，把本该从“实际摄像头”获取的视频流数据改为读取指定的“视频文件”对算法进行攻击。此类攻击属于设备底层技术篡改风险，攻击者通过数据劫持技术，对模型注入通过非法渠道采集或购买的真实数据，进而绕过 AI 模型对深度伪造和对抗攻击的防护，实现无感攻击，具有技术成熟、使用方法简单、隐蔽性强、检测难度大等特点。

1. 安全人脸安全 SDK

以人脸识别场景为例，百度自主研发的新一代人脸安全 SDK，如图 12 所示。通过设备软硬件信息、底层系统信息，人脸识别/活体认证行为信息计算出的安全因子，结合 AI 大数据风控技术，可有效识别通过技术手段对手机系统底层 ROM 进行篡改、注入、HOOK、或劫持摄像头等众多人脸识别场景面临的安全风险，如图 13 所示。

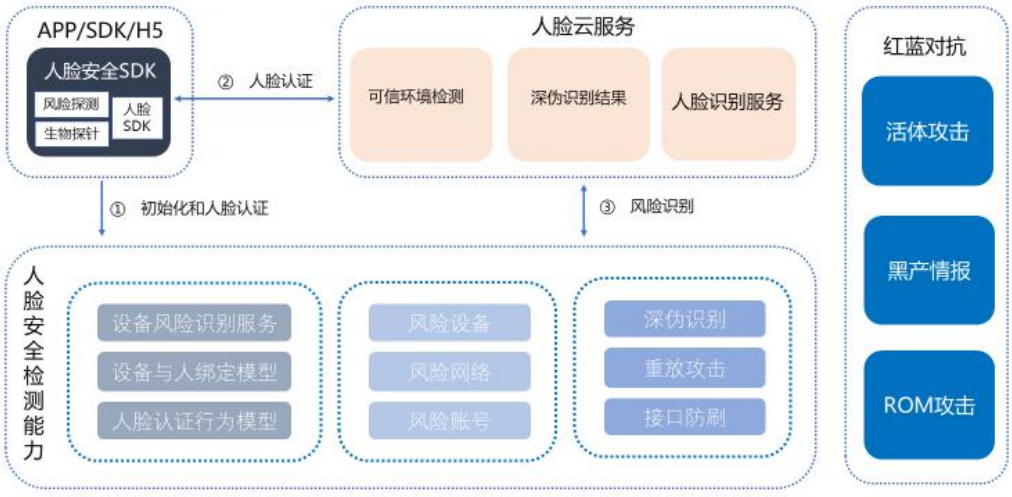

图 12 百度安全人脸安全 SDK 架构图

增强风控规则：增加对 ROOT、DEBUG、越狱等设备风险的多维拦截规则；

对活体图像进行加密回传：增加一机一密，以及白盒加解密，云端增加对数据解密的校验；

对 APP 进行加固防护：对活体控件进行全面加固防护，并对宿主 APP 进行部分加固，防止业务逻辑被代价被注入 HOOK；

图像追踪：增加图像的明文和隐藏数字签名，后端对数字签名的校验；

网络保护：升级网络通信链路协议，增加网络通讯双向校验，并使用安全 DNS 进行域名解析，防止通过中间人攻击方式对数据进行篡改；

行为检测：通过对用户人脸识别过程中的操作进行监控，精准识别非本人操作、设备姿态异常、认证环境异常等异常行为风险。

图 13 百度安全人脸安全检测

2. 人脸安全 SDK 应用：金融数据劫持风险检测

度小满金融 SDK 采用人脸识别作为信贷、理财、支付等核心场景的身份核验技术，面临多方监管安全合规要求、黑产攻击猖獗、以及传统人脸识别服务缺少数据劫持风险检测能力等三大问题，如图 14 所示。

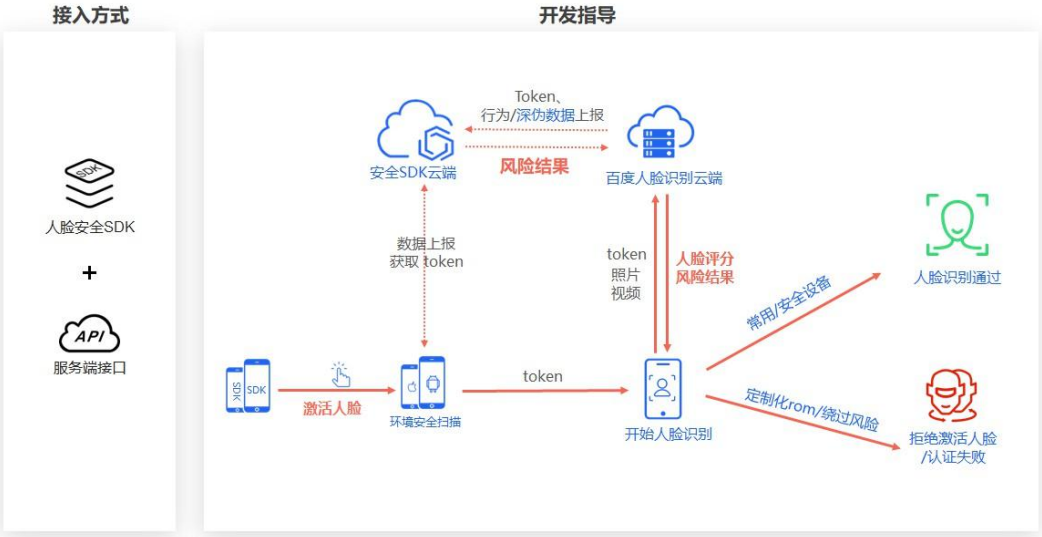

图 14 百度安全人脸安全检测

百度人脸安全方案的需要在客户端应用中集成人脸安全 SDK，云端调用人脸安全服务接口获取风险等级，根据业务进行精细化风险拦截。

百度人脸安全方案同时支持 SAAS 化和私有化两种接入方式，均可以实时返回业务环节风险等级。SAAS 化服务采用人脸安全 SDK+服务端 API 方式，成本低，接入快；私有化服务部署人脸安全的服务在客户的厂内环境，满足对数据流通的严格要求。

通过接入百度人脸安全方案，实现了对包括数据劫持在内的多种风险检测的能力，符合网信办、公安、银监会等监管要求，有效拦截并打击黑产犯罪行为。

（四） 360“框架缺陷”安全实践

框架缺陷是指机器学习框架代码中存在问题、缺陷，致使学习框架在训练或者推理时出现不满足预期的行为。在机器学习框架进行训练或者推理时，需要解析数据集或者加载模型文件，相关过程都有可能存在异常行为。框架缺陷类型包括内存非法访问越界、空指针引用、整数溢出等，危害性涉及信息泄露、拒绝服务、任意代码执行等。此外，在学习框架中，框架缺陷还包括精度丢失，使得 AI 服务无法满足预期要求。

1. 360 机器学习框架安全性评测工具 AIFater

云端机器学习框架中包含最基本的训练、推理功能，其中封装了负责大量计算操作的算子。不同计算型算子相互结合，形成了机器学习中的计算图，鉴于机器学习框架内部的复杂性，工具 AIFater 云端框架评测部分分为三个不同功能的模块：算子收集模块，算子测试模版库生成模块以及风险检测模块，如图 15 所示。

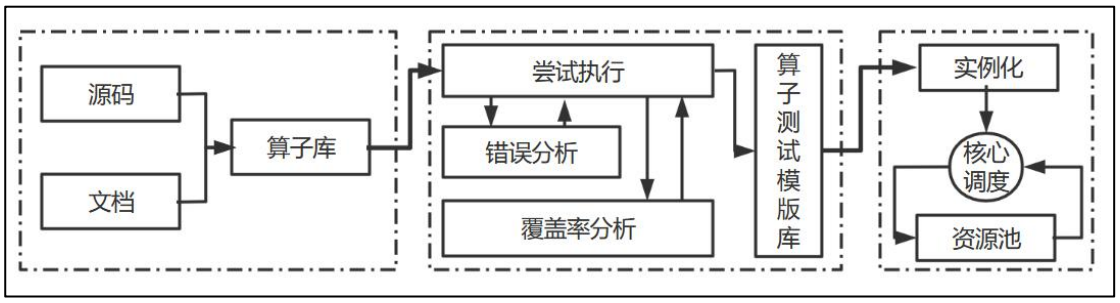

图 15 360AIFater 云端框架安全性评测部分

终端机器学习推理框架主要包含两部分功能：终端模型文件解析与机器学习推理。在解析阶段，框架通过模型文件中保存的网络拓扑信息，构建计算图；在推理阶段，框架通过用户输入数据，使用计算图进行推理，计算最终的结果。终端框架中封装了大量计算操作的算子实现，如卷积操作、池化操作、悉数操作等，不同算子由于自身功能不同，代码实现复杂度也随之不同。鉴于终端机器学习推理框架内部的复杂性，工具 AIFater 终端框架评测部分分为四个不同功能的模块：筛选模块、变异模块、特征提取模块和测试模块，如图 16 所示。

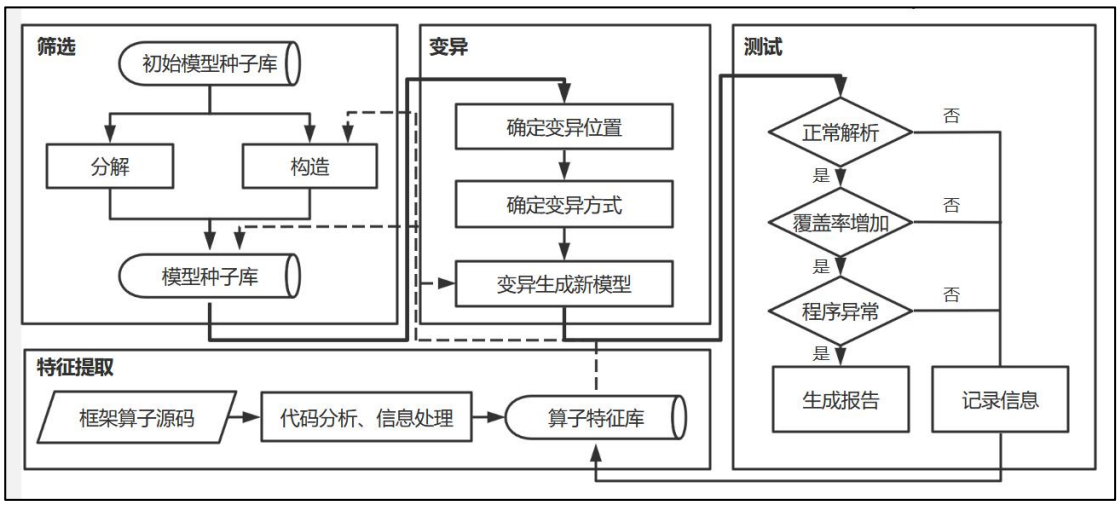

图 16 360AIFater 终端框架安全性评测部分

2. AIFater 应用：TensorFlow 框架安全性评测

机器学习框架可以帮助开发者高效、快速的构建网络模型进行训练和推理，而无需关心底层实现的细节。由于机器学习框架带来的高效、便捷的特性，越来越多的开发者加入到了 AI 应用研发的行列。谷歌推出的 TensorFlow 是目前最流行的机器学习框架之一，深受广大开发者的喜爱，其下载量已达上千万级，特别是在 2.0 版本发布后，其在简单和易用性方面有了大幅度的提升，极大的方便了开发者。然而，谷歌在研发 TensorFlow 的过程中，缺乏充分的安全性考虑，导致 TensorFlow 中存在诸多安全问题，可能会影响数以千万计的开发者以及用户。

AIFater 从算法实现、漏洞类型、编译优化等多角度持续开展对机器学习框架 TensorFlow 的安全风险研究，研制

了针对机器学习框架的安全评测平台 AIFater。AIFater 采用动、静结合的分析方法对不同语言实现（Python、C++、Go 等）中的不同漏洞类型进行全面系统的检测，检测过程着重关注从训练到推理、从数据到模型、从云端到终端过程中存在的安全风险。除了基本的安全性检测功能，AIFater 基于丰富的安全专家知识库，能够提供对 TensorFlow 定制化检测，如对计算图优化、图编译、分布并行等模块的安全性检测。

附件 2 行业安全防护实践案例

（一） 瑞莱“AI+金融”安全实践

1. 智能金融安全风险

近年来，人工智能在金融领域已经进入了规模化落地应用阶段，在广泛部署使用人工智能的同时对其安全性的保障尤为重要。一方面人工智能技术需要从海量数据中进行学习知识训练模型进行决策，但在使用数据的同时，需要对这些数据的隐私保护问题进行审慎考虑，防止数据被窃取滥用等隐患。另一方面，近年来研究发现人工智能算法模型本身具有鲁棒性较差等安全隐患，在金融领域针对于人工智能模型的新型攻击案件也在近几年来呈上升趋势，对于相关安全风险的防范也十分必要。因此，在当前人工智能的金融场景应用中，保障好数据安全和算法模型安全是保证人工智能安全可控应用的两个不可忽视的重要方面。

（1）模型安全问题

人工智能安全问题的一个重要表现方面是模型安全问题，其中较为突出的是人工智能模型鲁棒性不足，容易受到对抗样本攻击等的干扰。以人脸识别为例，其作为人工智能技术中的一项成熟技术在联网身份识别/核验、反欺

诈、线下支付、智慧网点等金融场景中被大量使用，但随之而来的非法破解牟利的黑灰产业也应运而生，金融机构面临着日益严峻的人脸识别攻击威胁和重大财产损失的风险。近年来针对于金融机构人脸识别认证环节的攻击方式层出不穷，由于人工智能模型不具备知识理解能力，对于输入进行预测时仅能通过设计者预设的判定策略做出决策，攻击者可从中找到漏洞实现破解。如攻击者通过制作与受害者脸部高度逼真的人物头模或人脸面具可以造成人脸识别系统的误识别；攻击者还可以恶意使用深度伪造技术，仅需一张受害者的人脸图片和开源的深度伪造算法或常用的表情制作 app 就可以“活化”人脸图片并欺骗人脸识别模型；另外近年来的人工智能相关研究还发现，人工智能算法模型本身具有特定的“缺陷”，易遭受对抗样本的攻击，在输入上恶意添加特定的干扰便可以导致算法模型识别出错。攻击者可通过此对抗攻击技术欺骗人脸识别模型，窃取他人身份验证标识。此类安全问题均会导致用户账户遭到不法分子的窃取利用、进而造成严重的财产损失等。

（2）数据安全问题

人工智能技术与数据资源息息相关，保障数据使用的安全性也尤为重要。2021 年 6 月，国家颁布了《中华人民共和国数据安全法》，提出了“数据是国家基础性战略资

源，数据资源是网络安全所保护的核心对象”，数据安全已上升到了国家战略高度。银行或各类金融机构在应用人工智能技术时需要大量数据资源。如在信贷分析应用场景中，银行或各类金融机构在构建或优化信贷分析模型时，通常会采用多家数据源联合建模，这需要多家银行或金融机构分别提供己方数据完成共同建模，一般情况下信贷分析模型所需要的数据涉及多个领域，且由于行业竞争、隐私安全和行政手续等问题，想要将各地、各机构的数据进行整合和使用是非常困难的。虽然各机构在提供己方数据时都已签署相关保密协议，但其约束力弱且仍存在法律风险，即使银行或金融机构选择建立中间方沙箱环境建模，也会存在拍照等等成本信息泄露的风险。所以，打通数据孤岛，使各家银行和金融机构间数据流通，在出台的政策和规范的引导下，高效的解决企业之间数据合作过程中的数据安全和隐私保护问题既是困难也是挑战。

2. 智能信贷隐私计算解决方案

北京瑞莱智慧科技有限公司是依托于清华大学人工智能研究院设立的人工智能企业，致力于提供安全可控的第三代人工智能，加速安全、可靠、可信的产业智能化升级，同时不断促进人工智能前沿学术成果的商业化落地。瑞莱智慧在人工智能安全领域有着领先世界的技术底蕴和

科研背景，该团队在顶级期刊中有百余篇论文，并在多个国际人工智能安全竞赛中取得冠军。在金融场景下，瑞莱智慧提供模型安全维度的人脸识别安全解决方案和数据安全维度的智能信贷隐私计算解决方案。

（1）人脸识别安全解决方案

瑞莱智慧的人脸识别安全解决方案包括两方面，一方面通过对人脸识别模型进行全方位的安全性测试来找出模型的安全漏洞和潜在隐患进而对模型安全性做出评估；另一方面，针对近年来的新型人工智能攻击，瑞莱智慧提供业界首发的人工智能安全防火墙来及时检测规避风险。

模型安全检测服务：使用打印的人脸图片、电子播放屏幕人脸内容和制作的人物头像和面具等、多种对抗攻击算法等对人脸识别系统进行破解的测试等，测试完成后对测试数据进行定量分析，进而发现被测系统中存在的安全漏洞和风险点，进而全方位地量化人脸识别系统进行全方位的安全性。

人工智能安全防火墙：针对人脸识别新型安全威胁，在人脸比对环节前部署 AI 安全防火墙，可以实现对人脸数据的安全性检测，当采集的人脸识别数据属于新型的对抗样本或深度伪造时，AI 安全防火墙能够及时反馈识别结果并指导业务系统加以拒绝，进而保障业务安全。

（2）智能信贷隐私计算应用解决方案

为高效的解决银行和金融机构间数据合作过程中的数据安全和隐私保护问题，可采用联邦学习的技术，保证各参与方在原始数据不出库的前提下，完成多方数据在加密机制下的参数交换并共同完成模型训练与预测。同时还需支持各方对进行分布式安全查询、统计的能力。除此之外还需要保证用户信息隐私的安全性，防止数据泄露的风险。

联邦学习：在整个模型训练过程中，不传输任何原始数据，充分保护用户信息隐私，确保数据协作合法合规。当前业界瑞莱智慧发布的隐私保护机器学习平台 RSC（RealSecure）可利用联邦学习技术实现机构间数据安全共用的需求，传统的联邦学习平台需要对不同算法进行联邦化，导致后续算法开发和更新时非常不便，RSC 平台提供底层数据流的视角来揭示机器学习算法和对应的分布式联邦学习算法的联系，并通过数据流图变换完成两者间的自动转换，可适配上层多种机器学习算法，进而简化后续算法更新和新算法集成的代价。

安全多方计算：此功能提供参与计算的各方进行分布式安全查询、统计的能力，同时也满足机构内部数据分析人员针对敏感或强监管数据进行查询和统计的诉求，在信任不足的情形下能获得数据合作计算的价值。安全计算底层主要借助秘密分享与同态加密算法实现，可实现在原始

数据、数据来源不暴露的情况下获取汇总统计结果。

匿踪查询技术：该技术能够提升数据交换的安全性，通常匿踪查询技术是基于半诚实对手的假设，通过隐藏被查询对象关键词或客户 ID 信息，使数据服务方提供匹配的查询结果同时无法获知具体对应的个体，从而更安全的服务于金融信贷场景。

3. 智能信贷隐私计算实践应用

（1）人脸识别模型安全落地实践案例

当前绝大部分银行可以使用人脸识别认证的方式来进行登录、修改个人信息和汇款等。因此，也出现了相关黑色产业利用了人脸识别系统的漏洞非法修改个人信息，非法汇款，导致了财产损失等问题。现已有实际落地的人脸识别安全检测业务和人工智能安全防火墙部署案例，切实的对新型人工智能安全风险实现了准确的检测和防护。

人脸识别安全检测业务主要包括对被测模型进行物理世界和数字世界的测评，此类业务模式已经落地于多家银行和金融机构，通过全面的测评，切实地发现了客户人脸识别模型的算法漏洞和安全隐患，并像客户提供测试报告，客户知悉被测人脸识别系统的相关漏洞后，以报告内容作为依据，对算法进行了针对性的加固和升级。

大多数银行或金融机构的人脸识别系统对于对抗样本

攻击和深度伪造攻击这类新型人工智能安全威胁表现不佳，且已经造成了财产和名誉的损失。针对于此类攻击方式，在原有的人脸识别流程中，部署加入人工智能安全防火墙，实现了对于含有对抗样本和深度伪造输入的检测和拒绝，识别时间达毫秒级，不影响用户体验，且识别准确率达到行方要求，部署示意图如图 17 所示。

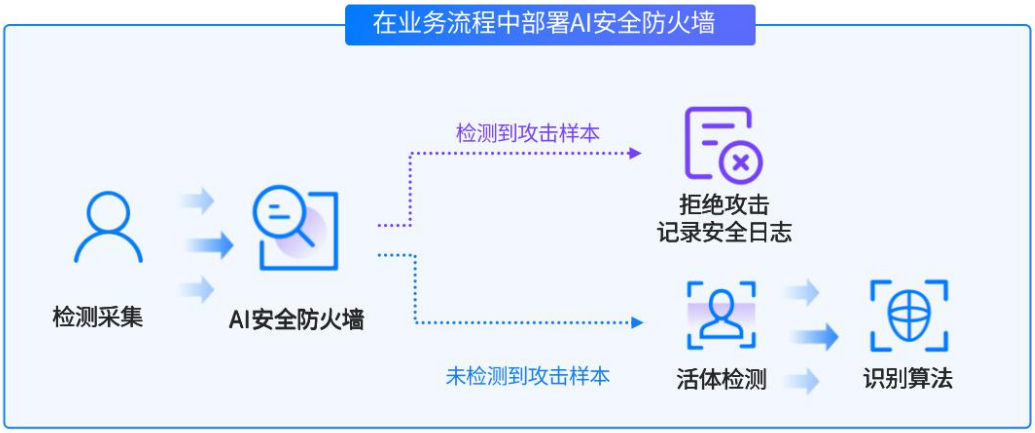

图 17 人工智能安全防火墙图

据某银行客户后续的反馈得知，该人脸识别安全检测服务切实找到了人脸识别模型的漏洞，并根据发现的漏洞进行加固和改进后，相关的安全事故和恶意行为大规模减小，冒用身份进行非法窃取的案件几乎杜绝。

（2）智能信贷数据安全落地实践案例

某股份行零售业务部在交易反欺诈工作开展的过程中，发现相关涉赌涉诈交易的甄别难度较大，主要原因是行内数据维度较单一，无法支撑相关分析与判断，另外在引入外部数据的过程中又涉及相关数据安全及隐私保护的

问题，故此希望通过隐私计算的模式安全合规的引入有效外部数据，构建针对涉赌涉炸交易甄别的反欺诈模型，辅助涉赌涉诈交易的甄别工作。

为了帮助银行有效解决当前反欺诈管理中遇到的困境，包括建模过程中行内数据特征维度不足及数据安全保护的问题，瑞莱智慧基于隐私计算技术，提供了“数据+平台+模型”的一体化解决方案，通过技术平台搭建、外部数据接入、联邦学习建模和赋能业务环节，通过在银行方及数据源方部署瑞莱智慧隐私保护计算平台 RSC（RealSecure），构建隐私保护计算节点，并接入 RSC 自带维度丰富的数据生态，然后基于行方样本数据进行联邦学习建模，模型训练完成后切线下效果验证通过后进行部署上线，嵌入到相关风控管理流程中辅助反欺诈工作开展。实现了银行与外部机构在反欺诈场景下的跨行业数据链接和更精准有效的欺诈甄别。

（二） 思必驰“AI+家居”安全实践

1. 智能家居安全风险

在智能变革的趋势下，智能穿戴产品也成为我们日常生活的必备，手机以及手机附属的智能耳机、智能手表、手环、智能眼镜等穿戴类带设备在我们的生活中十分常见。智能设备通常是一个相对封闭的系统，和网络发生不

间断的连接，其自主防护能力较弱，漏洞不容易被修复，形成了“稳定”的攻击源。而一旦爆发安全问题，将造成较大社会损失甚至人身安全威胁。传统的语音交互的应用一般需要联网进行识别和语义理解交互，类似苹果的 Siri 在离线情况下则无法进行语音交互，同时为了获得持续的语音体验，需要将本地数据传到云端进行模型的迭代更新，这就造成个人隐私数据的泄漏风险、APP 存在数据违规收集风险和个人设备存在误用风险。

（1）个人隐私数据的泄漏风险

人工智能时代，正让每个置身其中的人变得越来越透明。享受其便利的同时，公民的个人隐私也很容易被推向灰色地带。语音交互中涉及到的本地通讯录信息、手机本地的 APP 信息、家庭地址信息如果传到云端，很可能会对个人隐私信息进行暴露，一旦系统遭遇黑客攻击，被不法分子利用后产生不可挽救的损失。在语音技术的应用中，尤其是声纹识别技术的应用，用户的声纹特征属于个人隐私信息，一旦泄露，后果严重。

（2）APP 存在数据违规收集风险

经相关报告团队检测发现，超过九成的 App 都已具备隐私政策且内容丰富，但是其中超过半数 App 在用户首次登录时向用户默示隐私政策，导致隐私政策难以起到告知作用。63.1%的 App 通过“登录/注册即表示同意隐私政

策”的方式强制用户同意，且未提供拒绝选项，用户若想继续使用只能被动同意隐私政策，侵犯了用户自主选择权；16.9%的 App 在使用过程中仅展示隐私政策但未征询用户同意，违背制定隐私政策的初衷；15.4%的 App 提供了是否同意隐私政策的勾选框，但存在默认勾选问题，在用户不知情的情况下将风险让渡给用户，企图逃避自身责任。

（3）个人设备存在误用风险

语音交互可以缩短用户路径，但是如果允许他人可以随意交互的话，则对个人隐私的窥探增加风险，目前手机已经通过人脸识别实现快捷解锁，并可以通过活体检测避免照片等其他仿冒视频的恶意攻击，语音交互也需要对个人身份进行校验，并做好仿冒攻击方案。

2. 软硬一体化解决方案

思必驰产学研一体化的研发模式，与上海交通大学成立专属的联合研究实验室，并与苏州市人民政府联合成立“思必驰-上海交大苏州人工智能研究院”，配合思必驰内部研发团队，致力于智能语音语言及人机交互前沿技术的创新工作，同时不断强化技术的商业化应用及成果转化。

思必驰完全自主产权的全系列语音及语言交互技术，从感知到认知，形成人机智能交互的完整技术链条，核心技术包含语音识别、语音合成、语音识别++、语义理解、

智能对话等技术。另外，思必驰更关注人机对话式交互，多轮交互、打断纠错等技术处于国际领先水平，为产品提供专业深化的场景解决方案，为企业提供启发式对话的智能服务，同时开放 DUI 全链路智能对话定制平台，推进语音语言技术的应用规模化。

针对智能家居场景下的智能语音技术应用，思必驰提供 SDK 纯软件解决方案、麦克风阵列软硬一体化解决方案、AI 芯片方案等多种形式。此外，思必驰针对智能家居产品，更对接了丰富的后端内容服务。在智能家居产品的安全防护层面，思必驰也采取了更为有效的应对方案。

（1）个人隐私数据的泄漏解决方案

思必驰的智能语音技术方案，在用户从数据创建、存储、使用、共享直至销毁，提供全生命周期的安全保证，使用包括数据分级、数据加密、数据签名等措施，保障了数据的保密性、完整性、可用性、真实性、授权、认证和不可抵赖性。

数据分级：为所有用户和企业数据提供存储安全保护，按照数据敏感度和数据价值对数据进行等级划分，根据数据安全分级不同、对应不同的保护策略和要求，实现对用户和企业数据的分级保护。

数据加密：对用户密码、手机号、身份信息等敏感数

据加密存储，对敏感数据提供包括但不限于 AES 对称加密、RSA 非对称加密算法和信息摘要算法 SHA256 等标准算法以实现数据安全存储和保护，保障数据安全。

权限管理：所有用户和企业数据无授权任何人无权访问，访问用户隐私或敏感数据执行严格的控制权限申请流程。第三方应用访问与使用用户或企业数据，必需获取用户、企业明确可感知的授权。

动态脱敏：在数据输出到页面之前对敏感数据进行脱敏处理，支持手机号，身份信息以及其他自定义关键字段数据的脱敏保护。

安全审计：数据安全审计支持按照库、表、字段、访问源、数据库实例等多种维度进行审计规则设置，安全策略灵活且自由，可实现精细化监控；可根据不同数据级别进行个性化定制，精确控制数据库访问授权。审计内容包括登录时间、账户信息、SQL 语句类型、SQL 内容等信息，确保对数据库的操作行为在授权范围内。

（2）APP 存在数据违规收集解决方案

思必驰推出的智能语音解决方案中，配套的 APP 个人权限申请，遵循最小够用原则，仅收集使用业务功能必需的最少类型和数量的个人信息。数据存储是 App 运营过程中的关键环节，也是网络黑客攻击窃取数据的切入点，可

靠的数据存储为用户个人信息的正常使用提供重要保障。本方案配套在 APP 在不影响用户终端和服务正常使用的情况下，优先在用户个人终端内存储所收集的个人信息，并采取加密等技术措施确保用户数据即使泄露也难以被破解。

（3）个人设备误用解决方案

可以通过两种技术解决个人设备误用问题，一种是声纹仿冒攻击技术，通过声纹 ID 检测是否本人说话，并通过仿冒攻击技术避免录音与合成音攻击，加强智能电子设备的私人属性，防止个人数据被窥探。另一种是骨导人声检测技术，通过增加加速度或者骨震动传感器，可以通过采集骨震动信号判断本人是否在说话；当然缺点是播放音乐也会引起骨震动，这就需要我们对人声进行更好的提取。

3. 智能终端产品实践应用

思必驰为智能终端产品，尤其是智能家居类家电产品，提供全链路对话技术，完成从信号处理、语音识别、语义理解到智能对话的整个闭环，提供 SDK 纯软件解决方案、麦克风阵列软硬一体化解决方案、AI 芯片方案等多种形式。此外，思必驰针对智能家居产品，更对接了丰富的后端内容服务，包括资讯、音乐、电台、天气、购物、票务等，为用户打造一体化的交互服务体验，成为智能硬件

时代信息的重要入口。

思必驰的语音技术方案应用在了多款小米电视产品中，例如小米电视 6 OLED 和 ES 系列均搭载思必驰线性四麦前端信号处理、语音唤醒等技术，支持远场语音控制；小米电视大师系列 82"和 65”OLED 采用了思必驰前端信号处理算法与语音唤醒技术，并搭载思必驰 AMAEC 多通道回声消除技术，在全景杜比音效下也能实现流畅自然智能语音交互。语音技术的应用，让用户无需遥控器，不管是内容点播、音量调节，或者是查天气，智能家居联动…喊一声“小爱同学”，这台电视都可以为用户解决。

思必驰推出了多款麦克风阵列解决方案，四麦阵列在智能电视产品中有较多落地。思必驰四麦阵列解决方案集成降噪、唤醒、识别等功能，可拾取半径 5 米内的声源，可以对接其他智能硬件设备，配合思必驰丰富的后端资源，满足用户人机交互的需求。

相比于近场语音交互，远场语音交互中存在着信噪比低，信干比（信号比干扰）低，信回比（信号比回声）低，多径干扰严重等显著特点。思必驰四麦阵列结合语音信号和噪声的统计特性以及多阵元收集到的信源相位信息，可实现回声消除、噪声抑制、混响去除、声源定位等功能，从而有效提升了目标语音质量。与思必驰后端模型

相配合，可以达到远场复杂场景下的高唤醒率，高识别率，低误唤醒和低延时的目的。

（三） 蚂蚁“AI+生活”安全实践

1. 智慧生活安全风险

AI 技术已经是许多业务系统的核心驱动力，如苹果 Siri、微软小冰都依赖智能语音识别模型，谷歌照片利用图像识别技术快速识别图像中的人、动物、风景和地点。然而近年来学术界和工业界针对 AI 应用系统的攻击案例此起彼伏，例如 Facebook 和 Google 掀起了反 DeepFake 浪潮。

蚂蚁集团内部很多业务场景中也依赖于 AI 模型与系统，比如风控模型、人脸识别模型、智能客服模型等。AI 系统的防御与攻击也是一个不断演变的攻防对抗过程，攻击者会不断更新攻击手法来突破防御系统，尤其是以黑产为代表的攻击者，会不断探测 AI 系统的漏洞，开发新的攻击工具，降低成本来突破 AI 系统，获得更高的收益。

2. 多维对抗解决方案

为了保护 AI 模型的安全性，蚂蚁集团从多个角度来应对与攻击者之间日益焦灼的对抗战役。一个比较有效的战

略是从正面战场和敌后战场这两个战场来防御攻击者。在正面战场，采用多维对抗技术对抗攻击者，检测AI系统中的攻击案例；在敌后战场，应用AI模型安全开发生命周（AISDL）提升AI系统的安全性，事先发现AI系统中的漏洞然后修补，使攻击者的攻击手法时效。多维对抗是一种通过AI系统多个维度的数据异常探测，发现攻击的特征点，然后通过专家智能、机器智能进行攻击置信度排序和团伙挖掘等风险感知与定性的一体化架构。AI模型安全开发生命周期(AISDL)是从安全角度指导AI模型开发过程的管理模式。AISDL是一个安全保证的过程，它在AI模型开发的所有阶段都引入了安全和隐私的原则。具体来说，AI模型的生命周期包括模型设计、数据与与预训练训练模型准备、模型开发与训练、模型验证与测试、模型部署与上线、模型性能监控、模型下线这七个流程。AISDL通过安全指导这7个模型开发流程，保障模型在其全生命周期中的安全性。

3. 多维对抗实践应用

针对黑灰产、黑客、不法安全研究者的攻击，我们重点建设了4大环节的切面探测体系。具体分为设备侧、数据接口侧、模型侧和业务行为侧。基于这4大链路环节几十个切面探测点，我们进一步通过GroundTruth挖掘大脑

进行综合分析和挖掘，其包含策略模型、有监督模型、无监督模型、风险扩散模型和团伙挖掘模型等，将发现的单点异常提升到真正的威胁。这背后，我们的思考是在安全领域，模型一定是要充分利用专家经验的，如果直接拍脑袋进行特征抽取&建模，相信得到的结果里面一定会产生大量误报。因此，我们在切面探测层充分利用专家经验做广做深，而在威胁感知层，利用模型的复杂问题抽象能力和自动化能力去进一步提取威胁。在威胁感知和提取之后，我们进一步进行威胁研判，此时通过前面的步骤，人工需要处置的案件数量大大减少，可以进一步区分出攻击类别，进而从业务层面方便业务风险的定性。

传统的 SDL 流程能够有效保护应用的安全性，同样在 AI 模型生命周期中，可以细化为：模型设计->数据、预训练模型准备->模型开发与训练->模型验证与测试->模型发布->模型性能监控->模型下线。为了方便理解和实操，我们把流程进一步简化为三大模块：测评&加固、业务评审和风险治理。其中，业务评审环节，我们安全同学会参与到人脸识别的业务流程中，回答“在安全能力上，这个场景与这个模型是否匹配”的问题；风险治理环节，我们从实际的攻击溯源出发，让大家充分看见“威胁”，与业务方、风险策略同学共同确认场景的加固方案。

参考文献

[1]. 方滨兴.人工智能安全[M].北京：电子工业出版社，2020

[2]. ISO/IECTR24028：2020《信息技术人工智能人工智能的可信度概述》

[3]. CESA-2021-2-006《信息技术人工智能风险评估模型》团体标准（征求意见稿）

[4]. 2019-2020 人工智能发展报告[R].国家工业信息安全发展研究中心，2020 年 7 月

[5]. 人工智能标准化白皮书（2021）[R].中国电子技术标准化研究院，2021 年 7 月

[6]. 人工智能治理理论及系统的现状与趋势[J].计算机科学，2021 年 6 月

[7]. 人工智能安全白皮书（2020）[R].浙江大学-蚂蚁集团金融科技研究中心，数据安全与隐私保护实验室，2020 年 12 月

[8]. 人工智能安全框架（2020）[R].中国信息通信研究院安全研究所，2020 年 12 月

[9]. 睿思于前：AI 的安全和隐私保护[R].华为全球网络安全与用户隐私保护办公室，华为 2012 谢尔德实验室，2019 年 9 月

[10].人工智能安全标准化研究[J].保密科学技术,2019 年 9 月

[11].针对 ASR 系统的快速有目标自适应对抗攻击[J].西安电子科技大学学报,2021 年 1 月

[12].聚焦欧盟提案：构建统一人工智能治理体系.安全与保密，2021 年 7 月

[13].机器学习模型安全与隐私研究综述[J],软件学报，2021 年 1 月

[14].人工智能安全与测评[J]，人工智能前沿技术，2021 年

国家工业信息安全发展研究中心

地址：北京市石景山区鲁谷路 35 号

电话：010-88680530