CAICT 中国信通院

人工智能安全框架

（2020 年）

中国信息通信研究院安全研究所

2020 年 12 月

版权声明

牵头编写单位：中国信息通信研究院安全研究所

参与编写单位：北京瑞莱智慧科技有限公司

北京百度网讯科技有限公司

腾讯科技（深圳）有限公司

三六零安全科技股份有限公司

中国科学院信息工程研究所

编写组成员：魏亮、魏薇、景慧昀、彭志艺、张旭东、董胤蓬、王海棠、唐佳伟、邹权臣、张德岳、杨勇、刘水生、马多贺、徐世真、唐家渝、萧子豪、吴月升、黄超、唐梦云、陈昊、郝利民、孙振强、曹亮、刘昭、钱佳宇、周川、陈凯

前　言

人工智能作为引领新一轮科技革命和产业变革的战略性技术，正成为世界主要国家推动科技跨越式发展、实现产业优化升级、赢得全球竞争主动权的重要战略抓手。随着全球人工智能规模化建设和应用加速，人工智能基础设施、设计研发以及融合应用面临的安全风险日益凸显。世界主要国家纷纷通过制定人工智能伦理准则、完善法律法规和行业管理等方式开展人工智能安全治理。人工智能安全技术体系是人工智能安全治理的重要组成部分，是落实人工智能伦理规范和法律监管要求的重要支撑，是人工智能产业健康有序发展的重要保障。

人工智能安全框架是构建人工智能安全技术体系的重要指南，旨在为人工智能相关企业循序渐进提升安全能力、部署安全技术措施提供指导。在工业和信息化部网络安全管理局指导下，中国信息通信研究院联合瑞莱智慧、百度、腾讯、360、中科院信工所共同编制《人工智能安全框架（2020 年）》蓝皮书。本蓝皮书针对全球人工智能安全框架缺失问题，凝聚业界专家共识，聚焦当前人工智能突出安全风险，提出涵盖人工智能安全目标，人工智能安全分级能力，以及人工智能安全技术和管理体系的人工智能安全框架，期待为社会各方提升人工智能安全防护能力提供有益参考。

目 录

图 目 录

表 目 录

一、人工智能安全现状

(一)人工智能安全挑战

1.人工智能“基建化”加速，基础设施面临安全挑战

2020 年 5 月，我国《政府工作报告》提出以 5G、人工智能等为代表的新型基础设施建设政策，此举按下了人工智能国家战略推进的快进键。随后，25 省市发布“新基建”政策方案，累计投资 30 余万亿人民币，加快推动人工智能算力、算法和数据基础设施建设。在新基建推动催化下，人工智能技术将加快转变为像水、电一样的基础设施，向社会全行业全领域赋能。然而，人工智能基础设施却潜藏安全风险。以机器学习开源框架平台和预训练模型库为代表的算法基础设施因开发者蓄意破坏或代码实现不完善面临算法后门嵌入、代码安全漏洞等风险。2020 年 9 月，安全厂商 360 公开披露谷歌开源框架平台 TensorFlow 存在 24 个安全漏洞。开源数据集以及提供数据采集、清洗、标注等服务的人工智能基础数据设施面临训练数据不均衡、训练数据投毒、训练数据泄露等安全风险。2020 年，美国麻省理工学院的研究人员通过实验证实 CIFAR-100-LT、ImageNet-LT、SVHN-LT 等广泛应用的数据集存在严重不均衡问题。

2.人工智能“协同性”增强，设计研发安全风险突出

联邦学习、迁移学习等人工智能新技术的应用，促进跨机构间人工智能研发协作进一步增多。因遵循了不同目标和规范，使得人工智能设计研发阶段的安全风险更加复杂且难以检测发现。**一是人工智能**

算法自身存在技术脆弱性。当前，人工智能尚处于依托海量数据驱动知识学习的阶段，以深度神经网络为代表的人工智能算法仍存在弱鲁棒性、不可解释性、偏见歧视等尚未克服的技术局限。**二是人工智能新型安全攻击不断涌现。**近年来，对抗样本攻击、算法后门攻击、模型窃取攻击、模型反馈误导、数据逆向还原、成员推理攻击等破坏人工智能算法和数据机密性、完整性、可用性的新型安全攻击快速涌现，人工智能安全性获得全球学术界和工业界广泛关注。**三是算法设计实施有误产生非预期结果。**人工智能算法的设计和实施有可能无法实现设计者的预设目标，导致产生偏离预期的不可控行为。例如设计者为算法定义了错误的目标函数，导致算法在执行任务时对周围环境造成不良影响。

3. 人工智能“内嵌化”加深，应用失控风险危害显著

产业智能转型升级的内在驱动，不断推动人工智能深度内嵌于各行各业各环节中，真正实现物理世界变化实时映射于数字世界，以及数字世界演进优化带动物理世界发展的双向融合。然而，人工智能各行业应用带来的数字和物理世界双向融合，将促使人工智能在数字世界中的安全风险向物理世界和人类社会蔓延。**一是威胁物理环境安全。**应用于农业、化工、核工业等领域的智能系统非正常运行或遭受攻击，可能破坏土壤、海洋、大气等环境安全。**二是威胁人身财产安全。**自动驾驶、无人机、医疗机器人、智慧金融等智能系统的非正常运行将可能直接危害人类身体健康和财产安全。**三是威胁国家社会安全。**不法分子恶意利用基于人工智能的换脸换声技术伪造政治领袖和

公众人物的高逼真度新闻视频，可能引发民众骚乱甚至国内动乱，威胁国家安全。

（二）人工智能风险地图

与人工智能系统设计运营等全流程相结合，详尽剖析人工智能系统在各生命周期阶段面临的安全风险，将有助于分析定位人工智能安全风险来源，研究和部署针对性安全防御理论和技术。国际标准化组织（ISO）开展了《人工智能系统生命周期过程》标准项目，将人工智能系统全生命周期概括为初始、设计研发、检验验证、部署、运行监控、持续验证、重新评估、废弃八个阶段。基于 ISO 对于人工智能系统全生命周期的划分，项目组描绘出人工智能全生命周期安全风险地图，如图 1 所示。

初始阶段安全风险。初始阶段是指将想法转化为有形系统的过程，主要包括任务分析、需求定义、风险管理等过程。这个阶段的安全风险主要表现为对人工智能应用目标的设定有悖国家法律法规和社会伦理规范。

设计研发阶段安全风险。设计研发阶段是指完成可部署人工智能系统创建的过程，主要包括确定设计方法、定义系统框架、软件代码实现、风险管理等过程。这个阶段的安全风险主要表现为人工智能基础设施不完善、技术脆弱性以及设计研发有误等引发的安全风险。

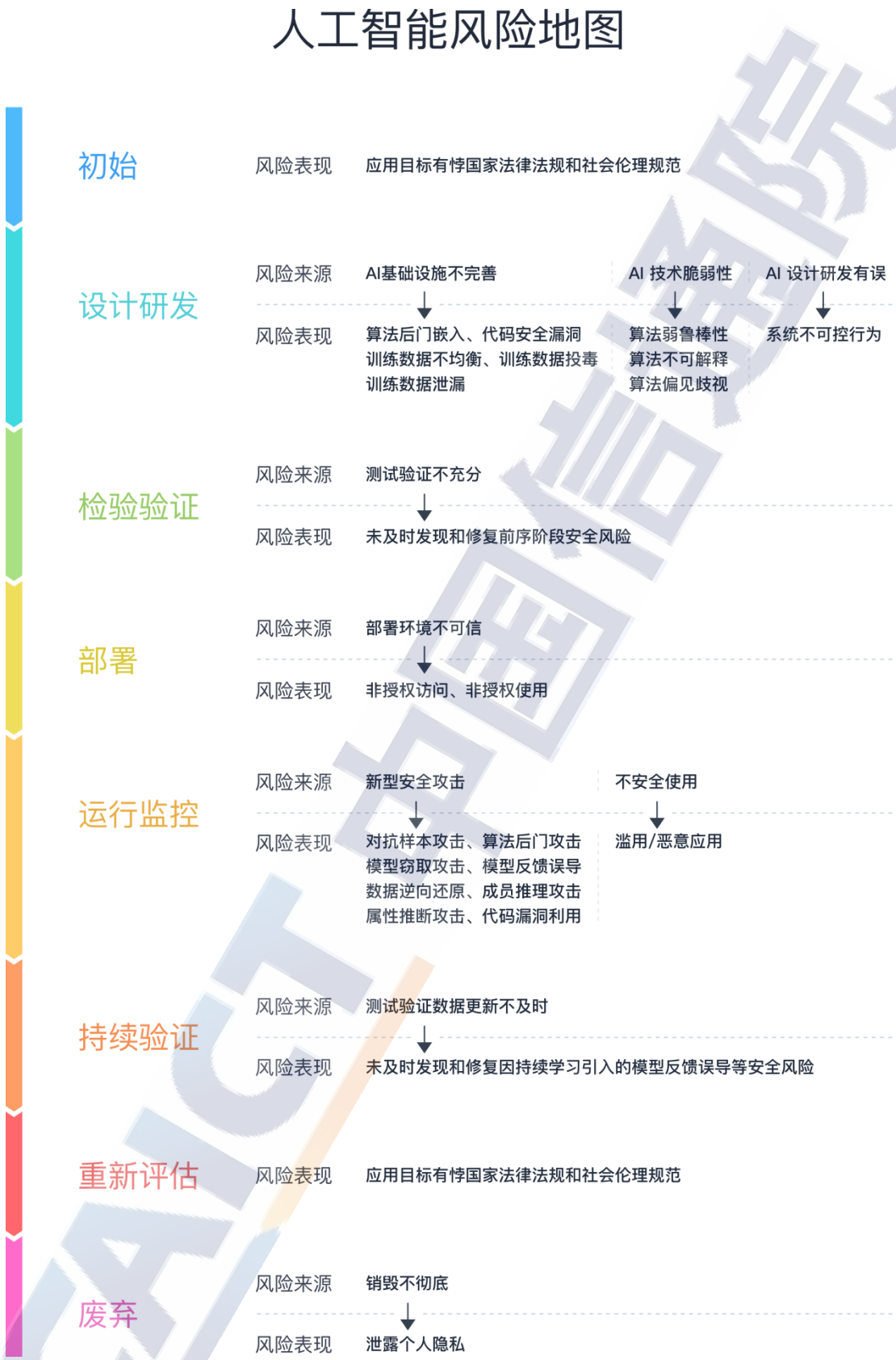

图 1 人工智能安全风险地图

检验验证阶段安全风险。检验验证阶段是指检查人工智能系统是否按照预期需求工作以及是否完全满足预定目标。这个阶段的安全风险主要表现为测试验证不充分，未及时发现和修复前序阶段的安全风险。

部署阶段安全风险。部署阶段是指在目标环境中安装和配置人工智能系统的过程。这个阶段的安全风险主要表现为人工智能系统部署的软硬件环境不可信，系统可能遭受非授权访问和非授权使用。

运行监控阶段安全风险。运行监控阶段，人工智能系统处于运行和可使用状态，主要包括运行监控、维护升级等过程。这个阶段的安全风险主要表现为恶意攻击者对人工智能系统发起的对抗样本、算法后门、模型窃取、模型反馈误导、数据逆向还原、成员推理、属性推断、代码漏洞利用等安全攻击，以及人工智能系统遭受滥用或恶意应用。

持续验证阶段安全风险。在持续验证阶段，对于开展持续学习的人工智能系统进行持续检验和验证。这个阶段的安全风险主要表现为测试验证数据更新不及时，未及时发现和修复因持续学习引入的模型反馈误导等安全风险。

重新评估阶段安全风险。当初始目标无法达到或者需要修改时，进入重新评估阶段。该阶段主要包括设计定义、需求定义、风险管理等过程。这个阶段主要涉及需求调整和重新定义，因而其安全风险与初始阶段的安全风险类似，即人工智能应用目标的设定有悖国家法律法规和社会伦理规范。

废弃阶段安全风险。在废弃阶段，废弃销毁使用目的不复存在或者有更好解决方法替换的人工智能系统，主要包括数据、算法模型以及系统整体的废弃销毁过程。这个阶段的安全风险主要表现为销毁不彻底，泄露个人隐私。

(三)人工智能安全技术现状

1.人工智能安全领域近年来论文数量增长迅速

近 10 年，人工智能安全风险和防御领域论文发表情况如图 2 所示。可以看出，自 2014 年谷歌研究人员首次证实深度神经网络面临对抗样本攻击威胁后，人工智能安全风险和防御领域论文数量迅速增长。

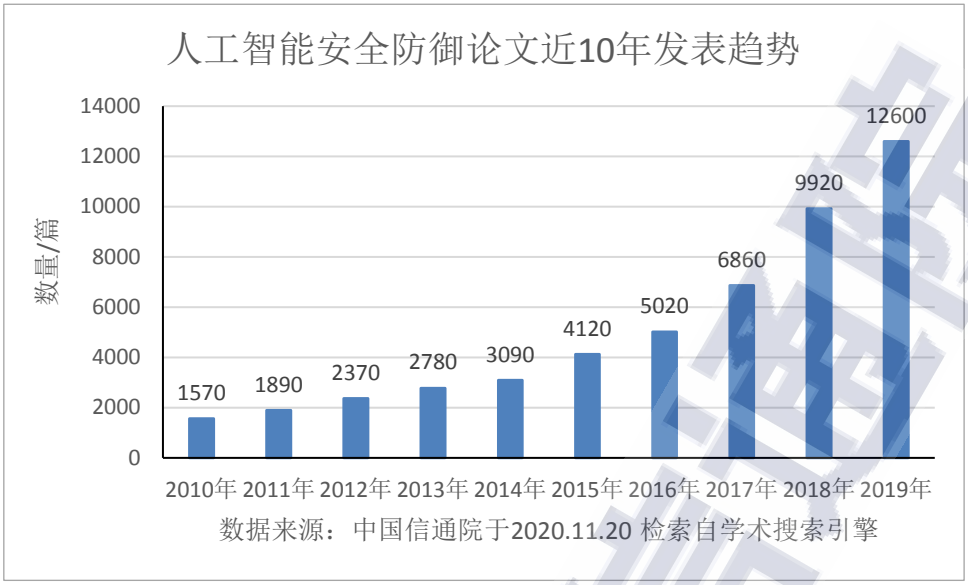

图 2 人工智能安全领域发表论文趋势

2. 人工智能安全热点技术方向

近年来，人工智能安全热点技术方向发表论文发表情况如图 3 所示。根据论文发表量可以看出，对抗样本攻击和防御是人工智能安全领域最受关注的研究方向。随后，数据投毒攻击和防御、模型可解释、算法后门攻击和防御这三个方向的论文发表量也均在 9000 篇以上，关注度较高。其次，联邦学习、差分隐私机器学习和深度伪造及检测近年来也逐渐成为受关注的技术方向。

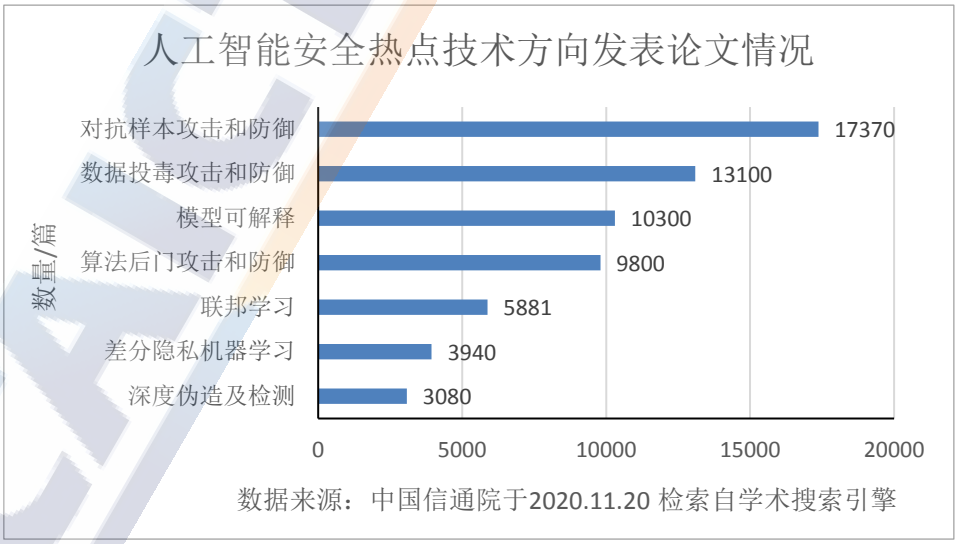

图 3 人工智能安全热点技术方向发表论文情况

随着人工智能技术应用愈加频繁，机器学习开源框架平台的安全性逐渐受到重视。机器学习开源框架平台安全漏洞挖掘修复也成为人工智能安全领域的热点研究方向。全球著名漏洞数据库 CVE 披露的典型机器学习开源框架平台安全漏洞数量逐渐增多，截至 2020 年 11 月 20 日的收录情况如图 4 所示。

图 4 CVE 收录典型机器学习开源框架平台安全漏洞数量

尽管人工智能安全热点技术方向大多是由美国研究人员首次提出，我国科研人员在相关领域已开展了大量创新性工作并取得了全球领先的研究成果。机器学习开源框架平台安全漏洞挖掘修复是由我国首次提出并贡献主要成果的人工智能安全热点技术方向。

表 1 人工智能安全热点技术方向提出国家及中国创新成果

序号	热点技术方向	提出年份	提出国家	中国创新成果
1	对抗样本攻击和防御	2014	美国谷歌公司研究人员首次证实针对深度神经网络的对抗样本攻击威胁[1]。	2017，清华大学朱军教授团队在有斯坦福、约翰霍普金斯等世界著名高校在内的 100 多支队伍参赛的 NIPS 2017 AI 对抗性攻防竞赛中，获得冠军。
2	训练数据投毒攻击和防御	2017	美国斯坦福大学首次证明了针对深度神经网络的对抗	2019 年，创新工场、南京大学等提出了一种高效的训练数据投毒方法，论文入选人工智能领域顶

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations (ICLR), 2014

			性投毒训练数据的存在[2]。	级国际会议 NIPS[3]。
3	模型可解释	2014	美国 纽约大学研究人员首次提出使用可视化对卷积神经网络进行解释的方法[4]。	2017 年，清华大学朱军教授团队开源珠算概率编程库大幅降低具有高可解释性贝叶斯深度学习算法的应用门槛。
4	算法后门攻击和防御	2013	美国 波多黎各理工大学首次提出神经网络木马攻击[5]。	2020 年，腾讯在第 19 届 XCon 大会上首次演示验证利用算法模型文件直接产生后门效果的攻击。
5	联邦学习	2016	美国 谷歌率先提出联邦学习概念[6]。	1. 2019 年，我国香港科技大学杨强教授提出了横向和纵向两种联邦学习框架； 2. 2019 年，微众银行推出了全球首个工业级联邦学习开源框架 FATE。
6	差分隐私机器学习	2016	美国 谷歌率先提出针对深度神经网络的差分隐私方法。	2020 年，第四范式研发的具有机器学习差分隐私保护能力的工业级平台先知（Sage）通过欧盟 GDPR 认证。
7	深度伪造及检测	2017	美国 名为 deepfakes 的用户在 Reddit 网站发布难辨真假的“假视频”。	2020 年，中国科技大学俞能海和张卫明教授团队在有全球 2265 支队伍参赛的 Kaggle 深度伪造检测挑战赛中脱颖而出，获得亚军。
8	机器学习开源框架平台安全漏洞挖掘修复	2017	中国 安全厂商 360 首次发现并披露机器学习开源框架平台供应链安全风险。	1. 腾讯发现首个 Tensorflow 安全漏洞； 2. 目前全球著名漏洞数据库 CVE 披露的 37 个 Tensorflow 漏洞中，24 个由中国安全厂商 360 发

[2] Koh P W, Liang P. Understanding black-box predictions via influence functions. arXiv preprint arXiv:1703.04730, 2017.

[3] Feng, J., Cai, Q.-Z., Zhou, Z.-H.: Learning to confuse: generating training time adversarial data with auto-encoder. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 11971 – 11981.

[4] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. In ECCV, 2014.

[5] Geigel A. Neural network trojan. Journal of Computer Security, 2013, 21(2): 191-232.

[6] Jakub Konecn ˇ y, H. Brendan McMahan, Felix X. Yu, ´ Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. In NIPS Workshop on Private Multi-Party Machine Learning, 2016

				现；截至 2020 年 11 月 20 日，360 累计发现框架平台依赖组件漏洞 100 余个。

3. 人工智能安全技术取得局部突破

人工智能安全热点技术方向中，联邦学习、差分隐私机器学习和深度伪造检测的商用步伐最快，已具有工业级产品并在部分领域开展试点应用。联邦学习方向，微众银行、字节跳动、京东数科等科技企业均推出了商用级联邦学习平台，并在保险定价、金融信贷、电商广告、智慧城市等领域开展试点商用。差分隐私机器学习方向，谷歌开源了差分隐私函数库 Differential Privacy Library，并已在谷歌地图、谷歌浏览器 Chrome 中开展实际应用。深度伪造检测方向，百度和瑞莱智慧推出了深度伪造检测服务平台，可向视频网站、网络论坛、新闻机构等提供人脸和人声伪造检测能力。

对抗样本攻击和防御技术方向处于由学术研究转化为商业应用的探索期，吸引了大量科技企业、科研院所和高校的关注。目前，已经涌现出 Cleverhans、Foolbox、ART、Advbox 等支持学术研究的开源工具，以及利用对抗样本攻击评测计算机视觉模型安全性的商用平台 RealSafe。阿里巴巴、腾讯、百度等科技企业通过举办人工智能对抗攻防大赛，积极发现针对人脸识别、图像分类、文本分析、目标检测等人工智能典型应用的有效对抗样本攻击和防御方法，为企业部署人工智能安全防护措施积累技术方案。

模型可解释技术为诊断发现“黑盒”人工智能算法模型缺陷提供可行路径，成为麻省理工大学、微软、谷歌、脸书、OpenAI 等全球

知名高校以及科技企业竞相布局的技术方向。麻省理工大学联合谷歌、伯克利等机构举办了 2018 年 NIPS“可解释性机器学习挑战赛”，有效推动了模型可解释技术发展。微软推出了 InterpretML 可解释开源工具包，不仅具有广义加性模型等自身具有可解释性的算法模型，而且提供对黑盒算法模型行为和预测结果进行解释的方法。谷歌在其云平台上推出了“可解释 AI”服务，旨在通过量化每个数据对模型决策的影响，帮助用户理解模型产生某项决策的原因。脸书开源的模型可解释库 Captum 以及 OpenAI 推出的神经元可视化工具 Microscope，能够帮助开发者更好地理解深度神经网络中不同神经元的行为和重要性。

（四）人工智能安全框架缺失

当前，随着数字经济和智能经济进阶发展，人工智能规模化建设和应用持续加速，人工智能安全风险日益凸显，并不断向物理世界和人类社会蔓延。保障人工智能应用安全可靠的需求日渐迫切。然而，现阶段企业主要聚焦于人工智能技术研发和产品运营，在人工智能安全方面投入相对较少、基础薄弱。而且，人工智能安全技术多处于学术研究和少量试验试用阶段，尚未形成适用于各类人工智能应用的成熟安全产品和服务体系。人工智能安全需求与企业安全投入不足以及人工智能安全产品服务欠缺之间的严重矛盾，成为制约人工智能产业健康发展的瓶颈问题。

人工智能安全框架，不仅能为企业提供提升人工智能安全能力的可参考路径，指导企业合理进行人工智能安全资源投入，并能为人工

智能安全技术研发提出方向建议，推动人工智能安全技术产品化和服务化。然而，目前全球尚未提出人工智能安全框架。而且，由于人工智能技术特点及安全风险与传统信息系统存在显著差异，现有网络安全框架并不适用于人工智能应用。因而，本蓝皮书聚焦人工智能安全风险，凝聚业界专家共识，构建人工智能安全框架。

二、人工智能安全框架

（一）设计思路

1. 框架范围

本人工智能安全框架聚焦于人工智能内生安全，即主要解决人工智能基础设施和人工智能设计研发面临的安全风险，以及因前两方面安全问题直接引发的人工智能应用行为决策失控安全风险。对于因滥用或者恶意使用人工智能应用而导致的物理世界和国家社会安全风险，主要由国家法律法规和行业监管政策对人工智能使用者予以规制。研发运营企业通过事前安全评估等内部管理机制可保障人工智能应用的目标符合国家法律法规和行业监管政策要求。

2. 设计原则

企业保障其设计研发和运营使用的人工智能应用安全是人工智能安全治理的基石。人工智能安全框架应为企业不断提升人工智能安全能力提供可遵循的迭代演进路径。为此，人工智能安全框架应具有较强的实用性、整体性和前瞻性。

实用性：当前世界主要国家和头部科技企业发布了人工智能伦理准则，提出了用于规范企业研发设计和运营使用人工智能行为的基本原则，但是并未提供落实相关原则的具体可实施方式。安全框架应对企业搭建人工智能安全防护体系和配置安全资源提供可操作的指导建议。

整体性：人工智能应用是集业务、算法、数据于一体的有机整体，并且其经历了初始分析、设计研发、检验验证、部署、运营监控等多个生命周期阶段。安全框架应面向人工智能应用的所有关键组件，涵盖人工智能应用的全生命周期，提出全面且有针对性的安全防护建议。

前瞻性：当前人工智能安全攻防理论和技术均在快速演变过程中。应当不仅局限于现阶段人工智能安全风险及防御技术，而是着眼于实现保障人工智能内生安全这一根本目标，提出能有效应对新的安全风险并兼容新的防御理论及技术的人工智能安全框架。

3. 核心要素

基于人工智能安全框架应遵循的实用性、前瞻性和整体性原则，从以下三个方面构建人工智能安全框架。

第一，明确人工智能安全目标是前提。目标的确定是一个根本问题，为人工智能安全防护工作的实施指明方向。本框架通过全面分析人工智能应用面临的安全风险，提出人工智能安全目标。

第二，构建人工智能安全能力是关键。为实现人工智能安全目标，本框架以建设人工智能安全能力为导向，参考网络安全滑动标尺，提出人工智能安全能力分级叠加演进模型。

第三，部署安全技术措施和落实安全管理是重要保障。为帮助人工智能应用研发运营企业有效形成和持续提升人工智能安全能力，提出了支撑实现人工智能安全能力的人工智能安全技术体系和管理体系。

综上所述，人工智能安全框架的构建包含安全目标、安全能力、安全技术和安全管理四个维度，从四个不同的层面指导企业开展人工智能安全防护工作。

（二）安全框架

人工智能安全框架包含安全目标、安全能力、安全技术和安全管理四个维度，如图 5 所示。这四个防护维度基于自顶向下、层层递进的方式指导企业构建人工智能安全防护体系。其中，设定合理安全目标是保障人工智能应用安全的起点和基础，安全能力是实现安全目标的有效保障，安全技术和安全管理是安全能力的支撑和体现。


人工智能安全框架
AI 产品和服务
人脸识别 自动驾驶
语音识别 …
需求
AI 安全目标
应用合法合规 功能可靠可控 数据安全可信
决策公平公正 行为可以解释 事件可以追溯
保障
AI 安全分级能力
反制进攻
针对人工智能恶意攻击者的合法反制安全能力
安全事件追溯
法律权益维护
威胁情报
获取和使用人工智能安全威胁情报，赋能人工智能安全系统、设备和人员的能力
情报管理 情报消费
情报生产
主动防御
引入和强化人工智能安全团队力量，实现动态、自适应、自生长的安全能力
持续安全监测 安全事件分析
安全防御响应 安全威胁预测
被动防御
在人工智能应用之外部署静态、被动式的安全能力
恶意行为发现 算法安全防护
数据安全防护
架构安全
用安全思维规划、设计、建设和使用人工智能应用，以提升其内生安全的能力
合规性评估 业务安全性保障
算法安全性增强 数据安全性提升 框架平台安全检测修复
支撑
AI 安全技术
业务安全
业务合规性评估
安全攻击检测
业务安全机制
恶意应用检测
算法安全
算法鲁棒性增强 算法公平性保障
算法可解释性提升 算法知识产权保护
算法安全评测
平台安全
漏洞挖掘修复
模型文件校验
框架平台安全部署
数据安全
数据隐私计算 数据追踪溯源
问题数据清洗 数据公平性增强
数据安全评测
AI 安全管理
企业层面
AI 安全组织架构 AI 安全人员能力
AI 安全制度流程
指导
行业层面
国家法律法规 行业管理政策
行业伦理准则 技术标准规范

图 5 人工智能安全框架

安全目标：通过系统分析人工智能面临的安全风险及其产生根源，从应用、功能、数据、决策、行为、事件六个方面提出安全需求和目标。

安全能力：按照安全能力建设难度逐级递增，以及安全资源投入产出比逐级递减的方式，参照网络安全滑动标尺模型，提出架构安全、被动防御、主动防御、威胁情报和反制进攻五级人工智能安全能力。前一级安全能力是构建后续级别安全能力的基础。其中第一级架构安全，旨在指导企业建立用安全思维规划、设计、建设和使用人工智能应用的能力。第二级被动防御，旨在指导企业在人工智能应用之外部署静态、被动式的安全能力。第三级主动防御，旨在指导企业强化人工智能安全团队，实现动态、自适应、自生长的安全能力。第四级威胁情报，旨在指导企业获取和使用人工智能安全威胁情报以赋能人工智能安全系统、设备和人员。第五级反制进攻，旨在指导企业建立针对人工智能恶意攻击者的合法反制安全能力。

安全技术：人工智能业务、人工智能算法、人工智能训练数据和机器学习框架平台是构建人工智能应用的四个核心组件，也是人工智能安全重点防护对象。因而，本框架针对业务、算法、数据和平台提出安全防护技术手段。

安全管理：从国家和行业人工智能安全法律法规、行业政策、伦理规范、技术标准等要求出发，提出企业在人工智能安全组织、人员和制度等方面的实施要求。

（三）框架分析

1. 安全目标

目前，欧盟、美国、中国等世界主要国家以及微软、谷歌等科技巨头均提出人工智能伦理准则。其中，合法性、可靠性、可控性、公平性、可追溯、隐私安全等安全目标成为人工智能伦理准则关注的重点。本安全框架在充分借鉴国内外人工智能伦理准则要求基础上，基于人工智能面临的安全风险和挑战，根据人工智能应用实际需要，提出以下六个方面安全目标。

应用合法合规：人工智能已在交通、医疗、领域展现出了强大的能力。滥用或恶意使用人工智能应用将会给物理世界和国家社会带来巨大的负面影响。因此，首先应确保人工智能系统应用目标符合国家法律法规和社会伦理规范要求。

功能可靠可控：人工智能技术正逐渐应用于智慧医疗、无人驾驶等安全关键性场景，人工智能的稳健可靠愈加重要。然而，对抗样本、算法后门等新型安全攻击方式，可通过修改运行时输入数据诱使人工智能应用产生非预期的错误输出。因而，应当确保人工智能系统各项功能在规定的运行条件和时间周期内始终产生预期的行为和结果，且一直处于人类操作员控制之下。

数据安全可信：数据是人工智能的基石，人工智能从数据中汲取知识的同时，也面临着数据泄露、数据偏见、数据投毒等诸多安全隐患。因而，应确保人工智能应用收集、使用、存储的数据不被窃取，

不会泄露用户隐私，且未被篡改，能够真实反映物理世界和人类社会情况。

决策公平公正：智能风控、智能招聘等人工智能应用正逐步辅助甚至替代人类进行关键决策。训练数据失衡、算法设计有误等原因可能导致人工智能应用产生带有偏见歧视的决策，损害国家社会公平正义。因而，应确保人工智能应用兼顾各类群体的特征信息，不会对特定人或群体做出带有歧视和偏见的决策。

行为可以解释：深度神经网络等人工智能算法的“不可解释性”，导致人们不仅无法解释算法做出某项决策的原因，也无法理解其内部运行原理和发现定位存在的问题。人工智能可解释性为诊断、发现、修复算法模型内在缺陷提供指导，是人工智能安全的基础。因而，应确保人工智能应用以人类可以理解的方式提供对其行为和结果合理性、准确性的解释。

事件可以追溯：人工智能算法的“不可解释性”，为人工智能安全事件的产生原因、行为主体等溯源要素分析带来挑战，传统安全审计方法无法胜任。因而，人工智能应用应根据业务场景量体裁衣，完善追溯体系，部署确保提供对安全事件产生原因、发生环节、行为主体等进行追踪溯源的技术措施。

2. 安全能力

当前，企业主要聚焦于人工智能应用的技术研发和商业运营，在人工智能安全方面的投入少、基础薄弱，无法短期内一蹴而就完成全部安全能力建设。而且，人工智能安全尚属于前沿创新领域，系统性

消减人工智能应用安全风险仍有待安全理论和技术的不断突破。为有效指引企业充分运用成熟安全技术循序渐进提升人工智能安全能力，本安全框架提出分级的人工智能安全能力模型。

确保人工智能应用内生安全，有效防范人工智能新型安全风险是本人工智能安全框架的主要范围和核心目标。网络安全滑动标尺是[7]详细探讨相关组织在提升传统信息系统内生安全、有效防御安全风险方面可实施的技术和管理措施的安全能力模型。网络安全滑动标尺模型共有五级，分别为：架构安全，被动防御，主动防御，威胁情报和反制进攻。各级安全能力之间具有连续性，后一级安全能力是前序级别安全能力的提升和扩展。

网络安全滑动标尺模型主要面向传统信息系统，每一级别规定的具体安全能力并不适用于技术特点及安全风险与传统信息系统存在显著差异的人工智能应用。因而，本框架在凝练借鉴网络安全滑动标尺模型核心思想的基础上，提出了人工智能安全能力分级叠加演进模型，系统规划了各级包含的人工智能安全能力。

（1）架构安全

架构安全指用安全思维规划、设计、建设和使用人工智能应用，以提升其内生安全的能力，主要包括以下五个方面。

合规性评估：在初始需求分析阶段，结合具体业务场景，评估人工智能应用的目标及方式是否符合国家法律法规、行业监管政策以及伦理规范。

[7] 网络安全滑动标尺模型：SANS 分析师罗伯特 梅里尔.李（Robert M. Lee）在 2015 年发布的网络安全框架。

业务安全性保障：在人工智能应用的业务层部署访问控制、安全隔离、安全熔断、安全冗余、安全监控等机制，保障在安全攻击等突发情况下人工智能应用仍能安全运行。例如，在自动驾驶汽车中部署安全分级回落机制，可保障危险情况下汽车控制权及时交还给人类。

算法安全性增强：通过改进算法训练方法、调整算法模型结构等方式，增强算法鲁棒性、可解释性和公平性等安全。例如可通过对抗训练、模型正则化等方式提升算法的鲁棒性。

数据安全性提升：通过数据隐私计算、问题数据清洗处理等方式，提升数据自身机密性、可用性。例如可通过差分隐私、同态加密、联邦学习等技术提升数据的机密性。

框架平台安全检测修复：对来自于第三方的预训练模型和机器学习开源框架平台进行安全检测，并对发现的安全问题及时修复，以提前感知风险，降低安全事件发生概率。例如 MindSpore[8]具有完善的漏洞管理流程，能够快速响应新提交的安全漏洞问题。

（2）被动防御

被动防御指针对人工智能的新型安全攻击，在人工智能应用之外部署静态、被动式的安全能力，主要包括以下三个方面。

恶意行为发现：通过分析提炼针对人工智能的新型安全攻击和恶意应用行为特征，实时对人工智能应用的外部访问、输入数据、行为决策等进行检测，及时发现对抗样本、模型窃取、算法后门、深度伪造等安全攻击和恶意应用行为。例如在人工智能应用外部增加检测组

[8] MindSpore：https://www.mindspore.cn/security

件或模型，利用正常样本和对抗样本在特征空间中的差异可检测对抗样本攻击。

算法安全防护：在人工智能算法模型外部部署安全防护组件，通过运用算法知识产权保护、问题数据重构、算法安全评测等措施，帮助人工智能应用有效抵御模型窃取、对抗样本等算法安全攻击。例如利用问题数据重构技术，在尽量保留原有图像语义的情况下，破坏攻击者恶意添加的扰动达到防御对抗样本攻击的目的。

数据安全防护：在人工智能应用外部部署安全防护组件，通过数据追踪溯源、数据安全评测等措施，帮助人工智能应用更有效抵御训练数据投毒、数据逆向还原、成员推理等数据安全攻击。例如利用数据安全标签技术，可及时发现被恶意篡改的数据达到防御训练数据投毒攻击的目的。

（3）主动防御

人工智能安全攻防技术正处于快速演化过程中，被动安全防御难以有效应对不断推陈出新的安全攻击手段。为弥补静态被动式防御的局限，主动防御旨在引入和强化人工智能安全团队力量，实现动态、自适应、自生长的安全能力，主要包括以下四个方面。

持续安全监测：能够在人工智能应用运行过程中，借助人工智能安全专家力量持续监测应用运行状况以及安全状态，给出应用当前安全风险级别，并对应用运行异常进行及时告警。

安全事件分析：在人工智能应用发生数据泄露、行为失控等安全事件时，通过引入人工智能安全专家力量及时分析研判事件的影响范围、严重程度、发生原因等。

安全防御响应：在安全事件发生时，人工智能安全专家综合利用各类安全防御技术及时对安全事件进行响应处置，并恢复人工智能应用的正常运行。

安全威胁预测：运用人工智能、大数据分析等技术，并结合人工智能安全专家的经验和洞察实现从历史数据中感知预测未知安全威胁。

（4）威胁情报

充分利用威胁情报信息将进一步提升和扩展主动防御效能。威胁情报是指获取和使用人工智能安全威胁情报，赋能人工智能安全系统、设备和人员的能力，主要包括以下三个方面。

情报管理：人工智能安全专家综合利用各类技术措施完成威胁情报的获取、分拣、分析、评级、分类等综合管理。

情报消费：人工智能安全专家综合运用威胁情报实现对未知威胁挖掘、系统防御策略更新以及安全设备能力增强。

情报产生：人工智能安全专家综合运用各类技术措施实现从各类公开数据资源中分析获取有关安全风险和威胁的知识。

（5）反制进攻

反制进攻指针对人工智能恶意攻击者的合法反制安全能力，主要包括以下两方面。

安全事件追溯：在安全事件发生时，确保所发生的安全事件能够追溯到相关实体，支撑后续的法律权益维护。

法律权益维护：出于自卫的目的，运用法律手段针对攻击者采取反击行为。

3. 安全技术

近期，人工智能安全领域已在算法鲁棒性增强、可解释性提升、数据隐私计算以及安全攻击检测和防御等方面取得了局部突破，可支撑实现人工智能安全分级能力模型中架构安全和被动防御初始两级的安全能力。然而，对于主动防御、威胁情报和反制进攻这三级所需的人工智能安全理论和技术仍待学术界和工业界进一步联合创新攻关。

（1）业务安全技术

业务安全技术指在人工智能业务层部署的安全防御技术。业务安全技术主要包括业务合规性评估、安全攻击检测、业务安全机制、恶意应用检测四个方面。

（2）算法安全技术

算法安全技术指针对人工智能算法部署的安全防御技术。算法安全技术主要包括算法鲁棒性增强、算法公平性保障、算法可解释性提升、算法知识产权保护、算法安全评测五个方面。

（3）数据安全技术

数据安全技术指针对人工智能训练数据部署的安全防御技术。数据安全技术主要包括数据隐私计算、数据追踪溯源、问题数据清洗、数据公平性增强、数据安全评测五个方面。

（4）平台安全技术

平台安全技术指针对机器学习框架平台部署的安全防御技术。平台安全技术主要包括漏洞挖掘修复、模型文件校验、框架平台安全部署三个方面。

4. 安全管理

人工智能安全管理体系中，国家政府发挥着领导性作用，统领管理机构的设立、法律法规的研制、监管政策的制定、技术标准的研发等方面。企业应当在充分理解遵守践行国家人工智能安全管理规则的基础上，不断完善自身的人工智能安全管理组织机构、制度流程和人员能力。

（1）行业层面

国家法律法规：我国已在数据和算法安全等人工智能基础性法律，以及深度伪造、智慧金融等具体行业法律法规方面开展了系列立法工作。例如，《数据安全法》《个人信息保护法》已经完成草案编制并提请全国人大常委会审议，这两部法律为我国人工智能应用的数据和个人隐私保护提供了基本遵循。《电子商务法》《网络信息内容生态治理规定》分别对广告推送、新闻推荐领域中的智能推荐算法进行了规制。《民法典》《证券法》分别对防范深度伪造滥用和程序化证券交易提供系统提出了规范化要求。

行业管理政策：整体而言，我国对于人工智能立法较为谨慎，对于自动驾驶、深度伪造、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。例如，《智能网联汽车道路测试管理规范（试行）》对测试主体、测试车辆以及测试申请、审核和管理等方面进行了规定。《网络音视频信息服务管理规定》对于使用深度学习、虚拟现实等技术从事音视频制作的信息服务机构提出了安全评估、以显著方式标识、建立健全辟谣机制等系列要求。《关于规范金融机构资产管理业务的指导意见》对智能投顾资质、算法备案管理、监管程序化交易算法失效进行了规范。《人工智能辅助诊断技术管理规范（试行）》《人工智能辅助治疗技术管理规范（试行）》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。

行业伦理准则：我国将伦理准则作为保障人工智能安全可靠可控发展的重要措施。2019 年 6 月，我国新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》提出了包含和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理八项原则的人工智能治理框架。2019 年 8 月，中国人工智能产业发展联盟发布《人工智能行业自律公约》，从行业组织角度推动人工智能企业伦理自律。

技术标准规范：我国已提出人工智能标准体系[9]，并将基础安全，数据、算法和模型安全，技术和系统安全，安全管理和服务，安全测

[9] 2020 年 8 月，国家标准化管理委员会、中央网信办、国家发展改革委、科技部、工业和信息化部联合引发了《国家新一代人工智能标准体系建设指南》

试评估，产品和应用安全作为人工智能安全与隐私保护技术标准的重要发力方向。在数据、算法和模型安全方面，我国信息安全标准化技术委员会已开展《信息安全技术 机器学习算法安全评估规范》《人工智能数据安全技术体系》国家标准研制和标准项目研究工作。在安全测试评估方面，我国通信标准化协会已开展《人工智能产品、应用及服务供应基础安全能力评估方法》《人工智能服务平台数据安全要求和评估方法》《人脸识别系统对抗鲁棒性安全技术要求和检测方法》等行业标准研制工作。在产品和应用安全方面，我国已发布《信息安全技术 远程人脸识别系统技术要求》国家标准，并开展了《生物特征识别信息的保护要求》《人脸识别数据安全要求》《步态识别数据安全要求》《声纹识别数据安全要求》等国家标准研制工作。

（2）企业层面

人工智能安全组织架构：企业设立或指定相关部门负责人工智能安全管理以及执行工作，明确人工智能安全的岗位职责以及人员分配，并建立有效的工作考核机制。与此同时，建立虚拟协同机构，在业务部、研发部、法务部、人力部等部门指定人工智能安全专员，负责企业统一的人工智能安全管理策略制定以及相关流程在本部门的实施。此外，为保障企业人工智能安全政策和管理要求的贯彻执行，企业设立人工智能安全监督机构，负责定期对人工智能安全制度执行情况、技术工具执行有效性等开展监督检查。

人工智能安全人员能力：人员的能力是企业保障人工智能安全的核心。人工智能安全是新兴的跨学科复合型领域，增强和扩大人工智

能安全专业人才队伍，提升从业人员专业技能是企业亟需完成的关键任务。人工智能安全的人员能力主要包括，人工智能安全管理能力、人工智能安全运营能力、人工智能安全技术能力。其中，人工智能安全管理能力是指，能够依据我国法律法规和行业监管政策要求以及业务特性，制定企业人工智能安全策略，编制人工智能安全制度流程规范文件，实施人工智能安全应急指挥和跨部门管理协调的能力。人工智能安全运营能力是指，在企业内持续性落实人工智能安全相关制度和流程，利用相关技术工具对人工智能安全风险进行监测、识别、预警和处置的能力。人工智能安全技术能力是指，了解国内外前沿的人工智能安全技术及发展趋势，熟悉国内外主流的人工智能安全产品和工具，能准确判断当前企业所需的最佳技术工具，并实际应用。

人工智能安全制度流程：从企业层面整体考虑和设计人工智能安全制度体系。企业内部人工智能安全制度体系中一般包括人工智能安全总纲、人工智能安全管理制度和办法，以及面向人工智能应用各生命周期的安全操作流程规范。其中，人工智能安全总纲明确企业人工智能安全管理的目标、愿景、策略、基本原则和总的管理要求。人工智能安全管理制度和办法，是指人工智能应用全生命周期阶段的安全制度要求。面向人工智能应用各生命周期的安全操作流程规范，是对人工智能安全管理办法的详细解释和补充，以便执行者深入理解和执行。

三、人工智能安全技术实施

（一）业务安全

1. 业务合规性评估

自评估：是指设计研发和运营使用人工智能应用的企业对本组织人工智能应用进行的合规性评估。在国家和行业标准以及企业内部规范的指导下，自评估企业对人工智能应用的目标、使用方式等是否符合国家法律法规、行业监管政策以及伦理规范中人工智能安全要求进行评估。目前，微软、旷视等国内外科技企业相继发布人工智能伦理道德原则，并成立人工智能伦理委员会对本企业人工智能应用进行评估审查。

第三方评估：是指社会第三方评估机构依据国家和行业标准，对人工智能应用的合规性进行评估。第三方评估可独立开展，也可在自评估基础上，对自评估中发现的存疑事项实施评估。例如，欧盟委员会在 2020 年 2 月发布的《人工智能白皮书——通往卓越和信任的欧洲路径》中明确提出，政府部门或第三方机构应对高风险人工智能应用开展安全评估。

2. 安全攻击检测

对抗样本检测：通过训练分类器检测输入数据是否为添加恶意扰动的对抗样本，及时发现预警人工智能应用是否正遭受对抗样本攻击。对抗样本检测方法主要有以下两类。一是基于数据特征层差异检测对抗样本，该类方法通过建模正常输入数据和对抗样本在数据特征

层的差异来区分识别对抗样本。例如，中国上海科技大学和美国加利福尼亚大学研究人员联合提出了 MagNet[10]检测方法，通过逼近正常输入数据的流形来训练分类器模型以检测对抗样本。澳大利亚墨尔本大学、美国加利福尼亚大学、中国清华大学等多所高校研究人员联合提出了基于局部本征维数的检测方法[11]，利用对抗样本的局部本征维数值远大于正常样本的性质识别对抗样本。二是基于模型预测结果差异检测对抗样本，该类方法利用算法模型在对抗样本上的预测结果与在对抗样本经去除扰动后的数据上的预测结果存在较大差异来检测对抗样本。例如，美国弗吉尼亚大学研究人员提出了特征挤压方法[12]，首先对输入数据进行颜色位深度压缩和像素值空间平滑的特征挤压操作，然后使用算法模型对原始输入数据和经过特征挤压的数据进行预测，如果两者的预测结果差距很大则判定输入数据为对抗样本。

算法后门检测：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 IBM 公司和芝加哥大学的研究人员联合提出激活聚类检测方法[13]，该方法利用正常训练数据和嵌入后门触发器的投毒数据的神经元激活存在显著差异的特性，设计了检测器以识别嵌入后门触发器的

[10] Dongyu Meng and Hao Chen. MagNet: a Two-Pronged Defense against Adversarial Examples. In ACM Conference on Computer and Communications Security (CCS), 2017

[11] Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Houle, M. E., Song, D., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. International Conference on Learning Representations, 2018

[12] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks. Proceedings of. The Network and Distributed System Security Symposium. (NDSS), 2018

[13] Chen B, Carvalho W, Baracaldo N, et al. Detecting backdoor attacks on deep neural networks by activation clustering. arXiv preprint arXiv:1811.03728, 2018

投毒数据。然而使用该方法需要事先拥有投毒数据，在实际检测过程中难以实现。此外，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[14]，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。然而，该方法计算成本较大，且不适用于检测有多类后门目标或者后门触发器修改量较大的情况。

目前，对于成员推理、属性推断、数据逆向还原、模型窃取等安全攻击，由于缺乏相关攻击的大量公开可用攻击技术手段，目前针对这些安全攻击的检测技术方法研究尚未获得广泛关注。

3. 业务安全机制

业务访问控制：通过身份验证、访问次数和访问频率限制等措施避免用户非法访问或恶意频繁访问，可有效防御攻击者利用大量访问结果估计人工智能应用内部信息实施模型窃取、成员推理等攻击行为。例如，讯飞、商汤等人工智能开放平台实施了对用户调用语音识别算法的单日访问次数和每秒访问频率进行限定的安全防护措施。

业务安全隔离：通过解耦人工智能应用各业务功能模块之间的逻辑关系，对各业务功能模块进行隔离，保障在部分业务出现异常时，其余业务仍能正常运行，进而提升人工智能应用的整体安全性。例如，百度自动驾驶开放平台 Apollo 采用了基于隔离技术的安全架构，通

[14] Wang B, Yao Y, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks, In 2019 IEEE Symposium on Security and Privacy, 2019

过部署车辆安全防火墙，有效隔离了车内与车外的网络，车身与车机的网络，保证各系统网络边界的独立。

业务安全冗余：在人工智能应用的关键业务环节部署多个完成相同功能的人工智能算法模型，使得在单个算法模型出现错误时不会影响到人工智能应用的最终业务决策，提升整个应用的可靠性。例如，腾讯人工智能医疗影像开放平台觅影通过融合 B 超、磁共振等多模态影像识别算法实现乳腺肿瘤筛查。

业务安全熔断：通过预先设定的安全策略，支持人工智能应用在动态环境中依据输出结果的不同置信度调整决策机制。例如当输出结果的置信度低于设定阈值时，人工智能应用将决策权交回人类控制者或者执行基于固定规则的判断决策。例如，红旗旗下的 L4 级自动驾驶汽车 EV 提供了在无法正确处理的特殊情况下靠边停车的安全熔断机制以保障自动驾驶的安全性。

业务安全监控：部署实时监测系统，对人工智能应用的运行状况和安全状态进行监控，分析研判人工智能应用当前安全风险级别，并对人工智能应用运行异常进行及时告警。例如，达芬奇手术机器人内置有实时监控系统，可及时发现预警手术机器人的误操作，保证手术的安全性。

4. 恶意应用检测

深度伪造检测：通过真实内容与伪造内容之间的特征差异进行真伪内容检测。根据检测对象的不同，可以将现有深度伪造检测技术分为针对图像的深度伪造检测方法、针对视频的深度伪造检测方法和针

对音频的深度伪造检测方法这三类。针对图像的深度伪造主流检测方法为利用深度神经网络提取人脸特征进行真伪分类。针对视频的深度伪造检测方法主要分为两类，一类方法首先将视频解析为视频帧，再利用图像检测的方法来对视频帧进行真伪判断，最终综合多帧判断结果给出视频是否为伪造的最终标签。另一类方法主要利用深度神经网络学习视频帧的时序特性，有效捕捉视频前后帧不一致、人脸不稳定、面部动作不连贯等特征，并以此为依据进行视频的真伪鉴别。针对语音的深度伪造检测方法主要通过深度神经网络提取伪造音频与真实音频之间语速、声纹和频谱分布的特征差异进行真伪鉴别。

（二）算法安全

1. 算法鲁棒性增强

数据增强：通过模拟自然场景或对抗场景中可能出现的各类情况，支撑算法模型从数据中学习到相关特征提升算法鲁棒性，从而在各种场景下始终保持正常的性能水平。数据增强方法可用于提升算法自然鲁棒性。例如，可以通过旋转平移、添加自然噪声等模拟不同场景下的干扰数据，利用数据风格迁移生成不易收集的场景数据等方式生成训练数据提升模型鲁棒性。Dan Hendrycks 等人开源的 Corruption and Perturbation Robustness 提供了可用于训练测试的多种场景下的模拟数据。同时，数据增强方法也可以用于提升对抗鲁棒性。例如，可以生成预加固模型的对抗样本数据用于对抗训练提

升算法模型鲁棒性。Madry 等人[15]将对抗样本数据与真实数据混合后进行对抗训练可大幅提升模型应对对抗样本攻击的表现，进而提升模型鲁棒性。

鲁棒特征学习：指通过使模型学习到在自然场景中不易被干扰的特征或降低对易被干扰特征的依赖程度来增强算法模型鲁棒性水平。例如，可通过修改调整模型损失函数、削弱易被干扰特征与模型决策之间的相关性等方式提升鲁棒性。Wang 等人[16]结合灰度共生矩阵（GLCM）对纹理敏感的特性来使得算法模型学习到更加稳定的特征，从而能够在保持较好识别效果的情况下大幅提升算法模型鲁棒性。

模型随机化：指通过在模型运行过程中引入随机性使得攻击者无法获得准确信息来优化攻击，进而确保算法模型在遭受主动攻击情况下依然能够保持正常的性能水平。例如可通过输入数据随机化，模型参数随机化、模型输出随机化增加攻击者代价。例如 Jeremy 等人[17]通过在输入图像上多次添加高斯随机噪声并以这些结果的平均值作为最后计算结果即可获得更好的鲁棒性。Liu 等人[18]通过利用具有参数随机化特性的贝叶斯神经网络来对输入数据进行预测，同样能够在保持较好识别效果的情况下获得较高的鲁棒性。

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks, arXiv:1706.06083.2017.
[16] Haohan Wang, Zexue He, Zachary C. Lipton, Eric P. Xing. Learning Robust Representations by Projecting Superficial Statistics Out, arXiv:1903.06256.2019.
[17] Jeremy M Cohen, Elan Rosenfeld, J. Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing, arXiv:1902.02918.2019.
[18] Xuanqing Liu, Yao Li, Chongruo Wu, Cho-Jui Hsieh. Adv-BNN: Improved Adversarial Defense through Robust Bayesian Neural Network, arXiv:1810.01279.2018.

模型正则化：指通过增加模型约束使其损失函数更加平滑，从而降低攻击者找到算法漏洞的可能性，进而确保算法模型在遭受主动攻击情况下依然能够保持正常的性能水平。例如，可通过模型权重正则化、模型梯度正则化等降低算法漏洞出现概率来提升算法模型对抗鲁棒性。Yan 等人[19]通过将模型与决策边界的距离作为正则项加入到训练过程中，增加了模型的决策边界与输入样本的距离，进而提升了攻击者添加扰动使模型越过分类边界的难度，成功增强了模型的鲁棒性。Daniel 等人[20]通过减少雅可比矩阵的范数，使得模型决策面更加平缓，同样达到了增加了攻击难度进而提升模型鲁棒性的目标。

训练数据采样：通过从原始数据中抽取子集组成训练集、限制每个用户可以贡献的最大数据量等方式，确保特殊样本不在模型训练数据中占据过大比例，进而有效避免故意和非故意因素导致的训练数据失衡后模型识别性能出现大幅下降的情况。例如，可通过逆变换采样、拒绝采样、重要性采样、马尔科夫蒙特卡洛采样法等防止训练数据分布失衡，进而提升算法鲁棒性。Google Gmail 邮件安全研究人员发现在 2017 年至 2018 年间，发生了 4 次通过将大量垃圾邮件反馈为非垃圾邮件的方式企图让 Gmail 垃圾邮件过滤器失衡的攻击。如果有针对性限制采集自每个用户的用于训练模型的数据量，即可有效避免此类问题。

[19] Ziang Yan, Yiwen Guo, Changshui Zhang. Deep Defense: Training DNNs with Improved Adversarial Robustness, arXiv:1803.00404.2018.

[20] Daniel Jakubovitz, Raja Giryes. Improving DNN Robustness to Adversarial Attacks using Jacobian Regularization, arXiv:1803.08680.2018.

2. 算法公平性保障

算法公平性约束：通过将算法公平性表示为模型约束的形式，将其加入到模型优化过程中，即可使训练得到的模型满足公平性要求，从而使模型对于任意输入数据都能产生公平决策。Kusner 等人[21]提出了多种公平性约束方法，如运用与敏感属性无直接或间接关联的属性建模、通过可观测变量的非确定性因素进行建模等。Bose 等人[22]针对现有的图嵌入算法无法处理公平约束的问题，在确保学习表示与敏感属性不相关的条件下，通过引入对抗框架来对图嵌入进行公平性约束，使用复合框架去除掉更多的敏感信息。Google 开源的 TensorFlow Constrained Optimization 也提供了相应的功能可供开发者调用。

偏见歧视后处理：指通过修改预训练模型对于任何输入的预测结果，使其满足公平性的要求。例如，Kamiran 等人[23]提出可按照公平性要求在不过度影响原有模型的识别效果的情况下修改叶子节点的预测类别，使得修改后的模型能够达到实现定义的公平性要求。Hardt 等人[24]在考虑到敏感属性的情况下，对不公平类别的概率估计进行后处理，学习不同敏感属性下的不同决策阈值，并在决策时应用这些特定阈值来提升公平性。IBM 开源的 AI Fairness 360 封装了多种偏见歧视改进后处理算法供开发者调用。

---

[21] KUSNER M, LOFTUS J, RUSSEL C, et al. Counterfactual fairness. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, USA, 2017.

[22] BOSE A J, HAMILTON W. Compositional fairness constraints for graph embeddings. (2019-07-16)[2020-07-07] https://arxiv.org/abs/1905.10674, 2019.

[23] KAMIRAN F, CALDERS T. Classifying without discriminating. Proceedings of 2009 2nd International Conference on Computer, Control and Communication. Karachi, Pakistan, 2009.

[24] HARDT M, PRICE E, SREBRO N. Equality of opportunity in supervised learning. Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook, USA, 2016: 3315−3323.

3. 算法可解释性提升

模型自解释：指通过直接使用自身具备可解释性的模型，无需额外信息也能够理解模型的决策过程和决策依据。例如，在决策树模型中，每一棵决策树都由表示特征或者属性的内部节点和表示类别的叶子节点组成，树的每一个分支代表一种可能的决策结果。决策树中每一条从根节点到不同叶子节点的路径都代表着一条不同的决策规则，可以通过将优化完成的决策树转变成 if-then 逻辑判断的方式对模型进行解释。类似的，清华大学人工智能研究院开源了珠算算法库，使开发者更容易使用具有较强可解释性和优良识别效果的贝叶斯深度学习技术。

算法全局解释：指以人类可理解的方式从整体上解释模型背后的决策逻辑和内部工作机制。算法全局解释是一类重要的算法事后解释技术，有多种方法可实现。例如，Liu 等人[25]利用模型蒸馏技术将复杂模型学习的函数压缩为更小更快的模型降低模型复杂度，进而更好的对算法进行解释。Yang 等人[26]通过从受训神经网络中提取隐含单元、输出单元等单个单元层次上的规则来解释复杂模型决策逻辑。Simonyan 等人[27]通过在特定的层上找到神经元的首选输入最大化神经元激活，来帮助理解神经网络神经元捕获的特征，进而帮助使用者从语义、视觉上理解神经网络的内部工作逻辑。

[25] Liu Xuan, Wang Xiaoguang, Matwin S, Improving the interpretability of deep neural networks with knowledge distillation. Proc of the 18th IEEE Int Conf on Data Mininng Workshops, Pinscataway, NJ: IEEEm, 2018: 905-912

[26] Yang Chengliang, Rangarajan A, Ranka S, Global mdel interpretation via recursive partitioning. Proc of the 4th IEEE Int Conf on Data Science and Systems, Piscataway, NJ: IEEE, 2018: 1563-1570

[27] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps C J J. arXiv preprint arXiv: 1312.6034, 2013

算法局部解释：指通过分析输入样本的每一维特征对模型的最终决策结果影响程度来帮助理解模型针对每一个特定输入样本的决策过程和决策依据。算法局部解释可以利用多种方法实现，例如 Saltelli 等人[28]提出的敏感性分析方法可以在给定的一组假设下，从定量分析的角度研究相关自变量发生变化后对某一特定的因变量影响程度进而对算法决策逻辑作出解释。Simonyan 等人[29]提出可以利用反向传播算法计算模型的输出相对于输入图像的梯度来求解该输入图像所对应的分类显著图进而推导输入数据特征重要性来解释算法决策逻辑。

4. 算法知识产权保护

模型水印：通过在训练时将水印嵌入模型文件，避免模型遭到窃取导致知识产权流失。Zhang 等人[30]提出了第一个用于保护图像处理模型的模型水印框架。他们发现攻击者使用目标模型的输入/输出对训练一个代理模型时，隐藏的水印也会被学习到。因此可以利用这种机制在模型被窃取之后进行取证溯源。例如，可通过在训练模型时在少量训练数据上添加水印并修改标签为特定类别，训练完成后的模型如果遭到第三方窃取，可尝试利用带水印的数据测试第三方模型是否会将该数据分类到特定类别，以此即可判断被测模型是否为原模型。

[28] Saltelli A , Tarantola S , Campolongo F , et al. Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models [M]. New York: John Wiley & Sons, 2004

[29] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv: 1312.6034, 2013

[30] Jie Zhang, Dongdong Chen, Jing Liao, Han Fang, Weiming Zhang, Wenbo Zhou, Hao Cui, Nenghai Yu, Model Watermarking for Image Processing Networks. arXiv:2002.11088.2020.

5. 算法安全评测

鲁棒性评测：目前人工智能鲁棒性评测主要包括以下两个方面。一是自然鲁棒性评测，该类方法通过输入数据变换、收集和重放小概率异常数据等方法模拟正常场景下的异常突变，评价人工智能应用在正常环境下的自然鲁棒性。例如，百度推出人工智能模型鲁棒性检测服务[31]，可通过对输入图像添加高斯、均匀、椒盐等噪声，实施水平位移、垂直位移、亮度调节、对比度调节、运动模糊等变换，模拟雾化、霜降等效果，实现对图像分类、物体识别等机器视觉算法的鲁棒性评测。二是对抗鲁棒性评测，该类方法通过使用对抗样本攻击等方法模拟恶意攻击者对输入数据的扰动，评价人工智能应用在恶意攻击情况下的鲁棒性。例如谷歌研究人员研发的 Cleverhans，IBM 推出的 ART，德国图宾根大学创建的 Foolbox，百度推出的 Advbox，瑞莱智慧发布的 RealSafe 等均具备对抗鲁棒性评测能力。

公平性评测：目前人工智能公平性评测方法主要有以下两类。一是基于静态数据集的公平性评测，该类方法通过分析对比人工智能应用在静态测试数据集中具有不同敏感属性值的数据组上的性能表现差异来评价人工智能应用公平性。例如，IBM 研发的 AI Fairness 360、微软提供的 Fairlearn、谷歌推出的 faimess-indicators 等均是基于静态数据集的公平性评测工具。二是基于动态模拟的公平性评测，该类方法通过强化学习等技术仿真模拟人工智能应用与运行环境的

[31] 百度人工智能模型鲁棒性评测服务：https://anquan.baidu.com/product/robustness

交互，用于评价受运行环境反馈严重影响的人工智能应用的长期公平性。谷歌推出的 ML-fairness-gym 即是基于动态模拟的公平性评测工具的代表。

可解释性评测：由于不同应用场景针对可解释的定义有较大区别等原因，目前进行可解释评测时，仍主要依赖人类主观观察判断人工智能应用的决策原理和原因是否透明易懂等定性方法，尚未有定量的可解释性评测方法和工具提出。缺乏可信的定量评价指标用于衡量和对比不同人工智能算法模型的可解释性，是该领域当前面临的突出瓶颈问题。

（三）数据安全

1. 数据隐私计算

安全多方计算：指基于多方数据协同完成计算目标的密码技术，实现除计算结果及其可推导出的信息之外不泄漏各方隐私数据。例如，可通过混淆电路技术只发送密文或随机数而不泄漏有效信息、在群体中合理分配秘密达成有效的秘密分享等方法来保护多方计算场景中的数据安全性。华控清交发布的 PrivPy 多方安全计算平台实现了支持通用计算类型、高性能、集群化和可扩展的解决方案。矩阵元发布的 JUGO 多方安全计算平台能够帮助数据提供方在保护数据安全隐私的前提下，实现数据的流程增值。

同态加密：是基于数学难题的计算复杂性理论的密码学技术，能够保障在对经过同态加密的数据进行计算的结果与对未加密原始数

据进行同一计算并加密处理的输出结果是一致的。例如，利用半同态加密算法，可保证对两个密文数字进行加减法操作后能够解密得到相应的明文加减法操作结果。而利用全同态加密算法不仅支持上述半同态加密算法，还支持乘法操作以实现更复杂的密文运算。微软开源的同态加密算法库 Microsoft Seal 封装了多个全同态加密算法，易于编译并可在不同环境中运行。Enviel 发布的零泄漏计算架构基于同态加密技术为公司和企业提供能抵御黑客威胁的加密数据使用平台。

零知识证明：指证明者能够在不向验证者提供任何有用信息的情况下，使验证者相信某个论断是正确的。零知识证明实质上是一种涉及两方或更多方的协议，即证明者向验证者证明并使其相信自己知道或拥有某一消息，但证明过程不能向验证者泄漏任何关于被证明消息的信息。零知识证明可以通过验证者与证明者之间的提问以及交互来进行，只要验证的次数足够即可获得证明，也可以通过由可信保密的第三方参与并进行证明。Ufile Chain 诚信档案联盟链平台运用零知识证明技术在保证个人信息隐私安全的同时，提供个人信息的认证、存储、流通、确权等服务。

差分隐私：通过在数据中添加干扰噪声的方式来避免攻击者分析数据集反向破解其中的数据与个体的对应关系，从而保护数据中的隐私信息。例如可通过应用拉普拉斯机制向确切的查询结果中加入随机噪声实现对数值型结果的保护，或应用指数机制在接收查询时以一定的概率值返回结果实现对离散型结果的保护。苹果公司在收集用户数据改善产品体验时已利用了差分隐私技术防止攻击者从中获取特定

个人信息。谷歌对外开源了差分隐私函数库，能够帮助开发者获取最终统计输出结果的同时不泄漏特定个人的信息。脸书开源的算法库 Opacus 通过引入差分私有随机梯度下降算法进一步优化了差分隐私的效率和安全性。

可信执行环境：指服务器和移动端的 CPU 中能够为数据和代码执行提供安全空间的一块区域。可信执行环境中运行的受信任应用程序可以访问设备主处理器和内存的全部功能，而硬件隔离保护这些组件不受主操作系统中运行的其他应用程序的影响。例如，Intel 提出的 SGX 机制，提供计算力、内存隔离能力。ARM 提出的 Secure EL2 可应用于移动端，能够使得多个安全操作系统共存，提供更细粒度的安全隔离机制。阿里云提出的 Link TEE 主要针对各种物联网设备提供可信执行环境安全框架。百度的 MesaTEE 能够应用于金融、自动驾驶、医疗等关键场景，提升业务运行时的安全性。

联邦学习：指在各参与方数据不出本地的情况下，通过加密机制下的参数交换方式进行数据联合训练建立共享机器学习模型的过程。针对不同的应用场景，联邦学习有多种模式可供选择。在用户特征维度重叠较多而用户重叠较少的情况下可采用横向联邦学习。在用户重叠较多而特征维度重叠较少的情况下可采用纵向联邦学习。在用户与用户特征维度重叠都较少的情况下采用联邦迁移学习方式。微众银行推出了能提供一站式联邦模型服务解决方案的 FATE，覆盖横向联邦学习、纵向联邦学习和联邦迁移学习。腾讯 T-Sec 联邦学习方案能够有效提升联合建模场景下的安全性。瑞莱智慧发布的 RealSecure 在

支持联邦学习核心特性的基础上，能够极大方便联邦学习算法开发和算法更新集成。

2. 数据追踪溯源

数据安全标签：将数据采集来源、采集时间、提供者、哈希值等有关原始数据的重要信息进行整合和加密处理，生成数据的安全标签。而后，通过数字水印等方式，将安全标签以不破坏源数据使用价值的方式隐藏在原始数据中，在溯源时即可通过提取安全标签，追踪到问题数据的提供者，并通过比对哈希值判断数据是否被篡改。天融信、安华金和等安全厂商推出的数据安全平台实现了数据标签功能。

区块链追踪溯源：将数据标识、采集来源、采集时间、提供者以及针对数据的每一次处理行为和处理者等数据溯源信息存储在区块链中，可实现对数据每一次处理行为的追溯。而且，区块链技术的去中心化、不可篡改以及可追溯等特点，可保障上链的溯源信息真实可靠。例如，阿里云、京东万象推出了区块链溯源服务，可向金融、电子商务等多个领域提供数据溯源服务。

3. 问题数据清洗

异常数据检测删除：通过分析异常数据及正常数据的差异，发现并删除潜在的异常数据。业务运行中常见需要检测的两类异常数据为投毒数据和对抗数据。例如 Brandon 等人[32]首先利用可能包含投毒数据的数据集训练模型，然后利用该模型对所有输入数据的特征进行聚类、奇异值分解以寻找到特征分布偏离正常值的数据，并过滤掉这些

[32] Brandon Tran, Jerry Li, Aleksander Madry, Spectral Signatures in Backdoor Attacks. arXiv:1811.00636.2018.

投毒数据。Reuben 等人[33]通过对抗样本偏离正常数据分布，核密度估计得到的概率密度较低等特性进行对抗样本检测。

问题数据重构：通过对输入数据重构处理，在保留原始样本语义的前提下，破坏攻击者添加在真实样本上的对抗扰动，从而防御对抗样本、算法后门等攻击。例如可通过图像压缩、图像总方差最小化，对抗样本去噪等方式最小化对抗样本与真实样本的距离，使得处理后的对抗样本接近于真实样本的像素分布，从而破坏增加的对抗扰动。例如瑞莱智慧发布的 RealSafe 人工智能安全平台提供了封装良好的 SDK 可以直接部署用于输入数据重构。

问题数据修复：通过分析异常数据与真实数据的差异，将异常数据修复为可用的正常数据。例如可通过传播临域信息至缺失位置实现纹理合成，运用基于图像块的搜索方法寻找与污损区域相似的目标区域等方式将污损缺失区域修复正常。北大、鹏城实验室等提出了一种能够大幅修复污损照片的新型算法 StructureFlow。百度 AI 开放平台发布的图像修复服务能够智能去除指定位置信息的对应物体，并能自动使用背景内容修复缺损图像信息。

4. 数据公平性增强

数据分布修正：指通过修改训练数据的分布使机器学习算法在分布均衡的数据集上训练从而实现算法模型对任意输入数据的预测结果均是公平、无歧视的。例如 Feldman 等人[34]对训练数据的每个属性

[33] Reuben Feinman, Ryan R. Curtin, Saurabh Shintre, Andrew B. GardnerDetecting Adversarial Samples from Artifacts, arXiv:1703.00410.2017

[34] FELDMAN M, FRIEDLER S A, MOELLER J, et al. Certifying and removing disparate impact[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2015: 259-268.

进行修改，使得基于给定敏感属性子集的边际分布都相等。IBM 开源的 AI Fairness 360 工具箱封装了多种公平性预处理算法可供开发者调用。

5. 数据安全评测

数据安全评测：目前人工智能数据安全评测主要包括以下两个方面。**一是数据合规性评测**，通过对人工智能应用中数据收集、使用、传输等全生命周期中的行为进行检查，评估评测其是否符合《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》《信息安全技术 个人信息安全规范》等法律法规和国家标准规范要求。**二是数据泄露安全性评测**，通过模拟数据窃取、成员推理攻击、数据逆向还原等方法模拟对人工智能应用的数据窃取行为，评价人工智能应用的数据安全性。

（四）平台安全

1. 漏洞挖掘修复

代码审计：通过静态代码审计技术可以检测机器学习开源框架平台代码中的安全漏洞及编码不规范等问题，及时发现开源框架平台中存在的安全风险。例如 CodeFlow、CodeQL 等自动化工具可以实现基于多种漏洞匹配规则的代码静态安全分析。

模糊测试：模糊测试是目前主流漏洞挖掘技术之一。通过对机器学习开源框架平台中文件解析、模型加载等模块进行模糊测试，可以提前发现并修复框架中的安全漏洞。例如，TensorFlow 框架中包含

了基于 libfuzzer 实现的模糊测试工具，可以实现对部分框架代码的模糊测试。此外，360 AI 安全研究院研制了面向云端机器学习框架的模糊测试工具 QSand 以及面向终端机器学习框架的模糊测试工具 FBFuzz，并使用这两项工具发现了 24 个 Tensorflow 安全漏洞。

安全响应机制：通过建立快速安全响应机制，可以借助白帽子、安全研究团队等社区力量发现安全问题，降低机器学习框架平台的安全风险。目前，一些主流机器学习框架平台厂商都具有相应的安全响应机制，例如谷歌为 Tensorflow 框架项目设置专门的邮箱及时接受安全研究人员提交的安全问题，并快速做出响应和反馈。此外，Tensorflow 项目还建立了单独的漏洞详情描述页面，用于对外公布最新漏洞详情，提醒广大开发者及时升级版本，防止漏洞造成严重危害。

2. 模型文件校验

模型文件校验：通过对模型文件格式、大小、参数范围、网络拓扑、节点名称、数据维度等关键信息进行检测校验，可以在模型文件加载前发现模型文件中存在的安全问题，防止恶意人工智能算法模型文件被加载。目前，多数主流的机器学习框架平台在模型解析、加载过程中采用了模型文件校验功能。例如，Tensoflow 的 lite 功能中专门提供了模型文件验证相关的 API，用于检查 Tensoflow lite 模型文件是否合法。旷世天元 Megengine 使用哈希字段对模型文件进行校验，在一定程度上可以防止模型文件被篡改。

3. 框架平台安全部署

可信环境部署：通过在可信环境中部署机器学习框架平台，可以增强机器学习框架平台运行环境的稳定性，隔断可能对框架平台造成的危害。特别是在无法保证机器学习框架平台输入安全的情况下，建议在沙箱环境中运行，并尽可能小的提供系统权限。例如 Tensorflow 中算法模型通常被编码成计算图的形式，模型参数可以决定计算图的行为，如果有恶意模型被加载，可能导致任意代码执行的严重后果。

（五）安全技术图谱

为便于企业有效部署实施人工智能业务、算法、数据和平台安全技术，项目组遵循国际标准化组织（ISO）《人工智能系统生命周期过程》标准对人工智能系统全生命周期的划分方式，描绘出各环节建议部署的人工智能安全技术，如图 6 所示。

人工智能安全技术图谱

	业务	算法	数据	平台
初始	业务合规性评估 自评估 第三方评估	–	–	–
设计研发	业务安全机制 业务访问控制 业务安全隔离 业务安全冗余	鲁棒性增强 数据增强　鲁棒特征学习 模型随机化　模型正则化 训练数据采样 知识产权保护 模型水印溯源 公平性保障 算法公平性约束 偏见歧视后处理 可解释性提升 模型自解释　算法全局解释 算法局部解释	数据隐私计算 多方安全计算　同态加密 零知识证明　差分隐私 联邦学习 数据追踪溯源 数据安全标签　区块链溯源 问题数据清洗 异常数据检测　问题数据重构 问题数据修复 数据公平性增强 数据分布修正	漏洞挖掘修复 代码审计 模糊测试 模型文件校验
检验验证	–	算法安全评测 公平性评测　鲁棒性评测 可解释性评测	数据安全评测 合规性评测　安全性评测	–
部署	可信执行环境	可信执行环境	可信执行环境	可信执行环境
运行监控	业务安全机制 业务安全熔断 业务安全监控 安全攻击检测 对抗样本检测 算法后门检测 恶意应用检测 深度伪造检测	–	–	–
持续验证	–	算法安全评测 公平性评测　鲁棒性评测 可解释性评测	数据安全评测 合规性评测　安全性评测	–
重新评估	合规性评估 自评估 第三方评估	–	–	–

图 6 人工智能安全技术图谱

四、人工智能重点应用安全防护实践

随着人工智能安全风险日益凸显，综合运用各类快速涌现的人工智能安全技术保障人工智能应用安全的需求日渐迫切。在关乎人类生命安全、财产安全以及国家社会安全的部分领域，已探索开展了人工智能应用安全防护工作，例如自动驾驶、智能信贷风控和深度伪造已成为开展人工智能安全防护较为领先的三个领域。为便于企业借鉴已有人工智能安全防护经验，本章详细介绍上述三个领域人工智能安全防护实践情况。

（一）自动驾驶安全防护

1. 安全风险

自动驾驶作为汽车智能化和网联化发展的高级形态，是人工智能与实体经济深度结合的典型代表，已成为各国竞相发力的重要方向。自动驾驶技术正重塑未来的出行方式，便利人们日常生活。目前，自动驾驶应用已在我国多地部署运营。2020 年 7 月，苏州发布了全球首条城市微循环无人小巴市民体验线路，落地了全国首个常态化运营的城市公开道路无人小巴项目。2020 年 10 月，百度自动驾驶出租车服务在北京全面开放，10 月 12 日单天呼单量突破 2600 单。

安全出行是自动驾驶行业的首要原则。自动驾驶汽车是由云端服务、传感器、计算单元、自动驾驶算法、底盘动力系统等构成的一套复杂的系统。由于其众多组件暴露出了大量攻击面，带来了严峻安全挑战。根据风险来源不同，自动驾驶安全风险可分为传统网络安全风

险以及人工智能安全风险两类。传统网络安全风险主要有云服务安全风险、计算环境破坏、车云网络通信安全风险和内部网络通信安全风险。传感器数据干扰和自动驾驶算法攻击是突出的人工智能安全风险。

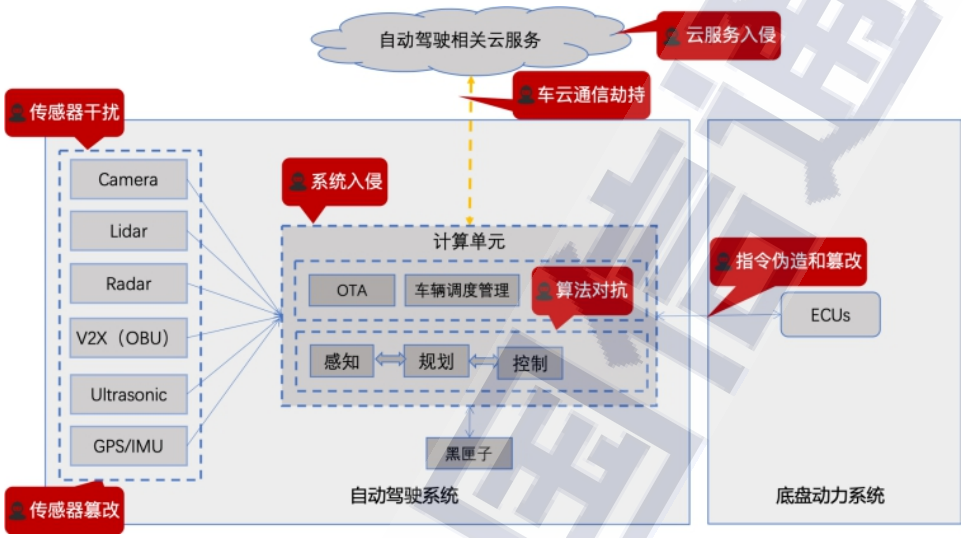

图 7 自动驾驶安全风险

一是传感器数据干扰风险。传感器是自动驾驶汽车的眼睛，通过观测周边环境，为自动驾驶核心算法提供外部真实世界的数据。攻击者可以通过光、电、磁等信号对摄像头、毫米波/超声波雷达、激光雷达、超视距传感器（V2X）、全球定位系统（GPS）等传感器进行干扰，从而使外部世界数据失真，威胁自动驾驶汽车的行驶安全。例如，在 2020 年国际安全极客大赛 GeekPwn 上，吴潍浠利用小型低成本的雷达干扰枪，实现了对自动驾驶汽车雷达系统干扰，使汽车无法正确识别前方障碍物并启动碰撞预警和自动紧急制动系统，进而直撞前方障碍物。

二是算法攻击安全风险。自动驾驶算法通过传感器采集的数据来感知周边世界，攻击者可以通过在现实世界中精心构造噪声和扰动来

对算法实施对抗样本攻击致感知异常，进而导致自动驾驶汽车行驶异常，威胁行车安全。与此同时，自动驾驶使用的人工智能算法需要大量标注数据进行模型训练，攻击者可通过构造错误的数据标注，导致算法模型训练结果失真，干扰算法模型预测的准确率，最终危及自动驾驶汽车安全。例如 2018 年国际顶级安全会议 BlackHat 上，百度安全团队演示了针对自动驾驶系统的物理域对抗样本攻击，即在目标车辆后方粘贴特殊制作的扰动图片可成功逃避自动驾驶汽车的车辆检测识别。

2. 安全防护

自动驾驶汽车的任一部分受到攻击都会危及整个系统的正常运行，单点防御策略很难保证自动驾驶整体安全性。因而需构建多层次的纵深防御技术体系，保障自动驾驶汽车安全行驶。百度针对自动驾驶安全防御提出并部署了包含云服务安全、外部通信安全、内部通信安全、计算环境安全、AI 算法安全、AI 业务安全六个方面安全防护技术。本节重点介绍与人工智能安全风险相关的安全防御技术。

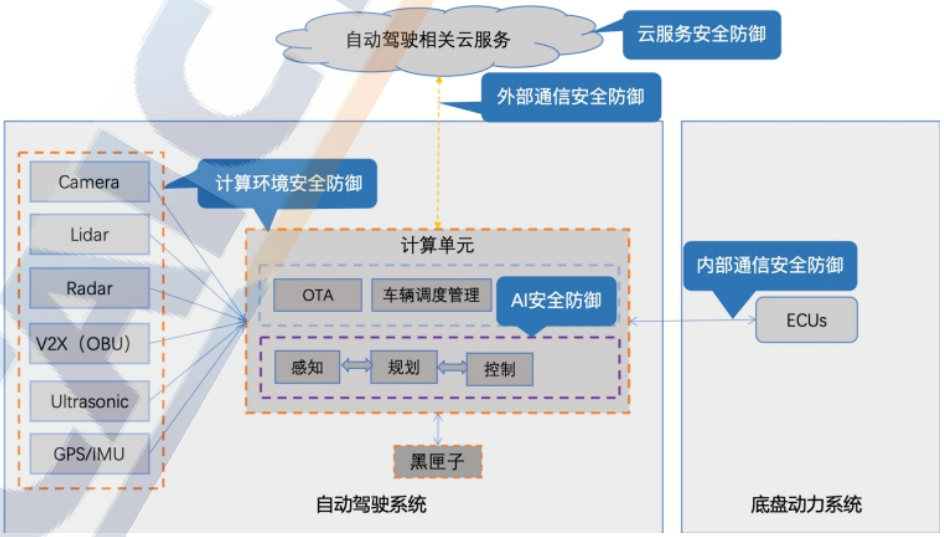

图 8 自动驾驶安全防护技术体系

(1) 计算环境安全

采用高质量且具有抗干扰能力的传感器，可有效应对传感器数据干扰攻击。一是安装具有电磁信号跳频、编码等特性的毫米波/超声波传感器和具有脉冲变频、光波编码等特性的激光雷达，提升抗干扰能力，增加攻击难度。例如，Baraja 公司推出的下一代传感平台使用随机调制连续波的激光雷达系统，可以阻挡环境光源的干扰。二是采用多源卫星接收机，提高传感器抗干扰能力。在遇到 GPS 信号伪造攻击时仍可以接收到其他卫星定位系统的信号。例如，百度自动驾驶系统 Apollo 推荐采用 Novatel PP7 多源接收机系统。三是对 V2X 消息进行签名验证，识别真实的 V2X 消息，丢弃虚假消息，确保真实可靠。目前，我国已开展两轮 V2X 安全消息一致性测试及示范活动，百度等自动驾驶头部公司已实际部署 V2X 安全机制。

(2) AI 算法安全防御

为了防御针对自动驾驶算法的对抗样本攻击，在算法模型训练数据集中引入对抗样本，增强模型的鲁棒性。与此同时利用模型加密、访问限制等技术，保护模型的机密性，防止算法模型泄露。例如，百度自动驾驶系统 Apollo 采用了包含模型安全验证、模型加固、对抗样本检测、模型鲁棒性形式化验证的一体化算法模型安全解决方案 AdvBox。

(3) AI 业务安全防御

综合运用降级预案、感知融合、关键数据运行记录等措施，实现自动驾驶的安全防御。

降级预案：当自动驾驶系统检测到攻击后，如果系统无法安全处理，则启动相应的熔断或降级预案，降低攻击造成的后果。例如，Waymo、百度等公司的自动驾驶车辆均设有一键停车按钮，在遭受攻击时可以降低车辆的安全风险。

多感知融合：通过多传感器感知算法融合，实现多感知算法识别结果相互交叉验证，增强自动驾驶系统抗传感器数据干扰能力。一是视觉感知和雷达感知的融合，即当一个传感器采集数据被干扰后，感知算法仍能使用其他传感器数据感知障碍物，确保自动驾驶汽车安全行驶。二是 V2X 消息、视觉感知和高精地图的融合，即当交通标识被篡改后，仍然可以通过 V2X 或者高精地图得到正确的交通标识信息。三是卫星定位、惯导定位和环境特征定位的融合，即在 GPS 信号受到干扰时，仍可通过惯导定位、环境特征定位感知车辆的正确位置。例如，自动驾驶公司 DeepScale、视觉方案供应商 Mobileye 均专注于传感器融合解决方案。大疆推出的 Mavic Air 通过多感知融合技术达到出色的定位效果。

关键运行数据记录：妥善保存自动驾驶汽车行驶过程中产生的关键数据，例如决策控制指令、车辆状态信息等，有助于发生交通事故时还原事故现场，定位事故原因，判定责任主体。在自动驾驶汽车中安装具有数据防删除、防篡改、防伪造能力的数据记录装置，有助于确保记录的关键数据安全可信。例如，配置自动驾驶系统的奥迪 A8 为车辆安装了“黑匣子”，用以专门记录驾驶权交接与机器警示驾驶

接管全记录，且会将记录数据保存六个月，以此为法律上界定事故责任提供技术支撑。

（二）智能信贷风控安全防护

1. 安全风险

银行信贷业务的风险管理核心在于构建风控模型，帮助银行有效识别客户信用风险以及欺诈行为。获取多领域多维度高质量用户数据对于建立识别客户风险的风控模型是至关重要的。然而利用多源数据进行风控建模面临诸多风险挑战。

一是数据泄漏风险。金融机构在构建或者优化信贷模型时通常会选择与数据提供方联合建模。由于传统联合建模方式至少需要一方数据出库，即使金融机构和数据提供方选择在中间方建立沙箱环境进行建模，也会存在拍照、截屏等低成本数据泄露安全风险。

二是数据孤岛风险。信贷模型的建模通常需要征信、电商交易、银行流水等多个领域的数据。然而，由于行业竞争、隐私安全、行政手续等壁垒问题，金融机构很难整合使用分散在各地、各机构的不同领域数据来训练更好的信贷模型。

三是安全合规风险。近年来，金融行业数据安全和个人信息保护的管理规范和技术标准日益完善。例如，《金融科技（FinTech）发展规划（2019-2021 年）》《移动金融基于声纹识别的安全应用技术规范》《个人金融信息保护技术规范》等对金融机构的数据收集、流转、使用行为提出了全面详细的安全要求，从业机构需快速适应新监管要求，确保业务安全性和合规性。

2. 安全防护

为打通数据孤岛、实现数据流动的价值，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，人工智能行业出现了多种数据安全防护方法，如蚂蚁集团、腾讯、瑞莱智慧等在智能信贷风控场景已利用安全多方计算、联邦学习、匿踪查询等技术打造数据安全共享基础设施，将计算移动到数据端，实现数据可用不可见，并最终帮助用户完成跨机构数据合作任务，驱动业务增长。主要的安全防护方法包括以下三方面：

（1）安全多方计算

安全多方计算提供参与计算的各方对敏感或强监管数据进行分布式安全查询、统计和复杂计算的能力，在信任不足的情形下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计算结果。例如，蚂蚁集团推出了蚂蚁链摩斯大规模多方安全计算商用平台，基于多方安全计算、区块链等技术，解决企业数据协同计算过程中的数据安全和隐私保护问题。蚂蚁链摩斯平台已在金融风控、联合营销等场景中进行运用。

（2）联邦学习

联邦学习保证在各参与方自有原始数据不出库前提下，两方或多方通过加密机制进行参数交换，实现人工智能算法模型协同训练与预测。金融机构在人工智能算法训练过程中，不对外传输任何原始数据，杜绝数据泄露隐患，充分保护用户隐私，确保数据协作合规性。例如

瑞莱智慧研发的隐私保护机器学习平台 RealSecure 在支持联邦学习核心特性的基础上，采用底层人工智能编译器架构，无需针对每个参与方编写对应的计算逻辑，可极大方便联邦学习算法开发以及后续新机器学习算法的不断更新和集成。目前瑞莱智慧已帮助多家机构解决数据合作过程中的数据安全风险和隐私泄露问题。

（3）匿踪查询

匿踪查询技术是基于半诚实对手的假设，利用不对称加密、不经意传输等密码学技术，通过隐藏被查询对象关键词或客户身份信息，使数据服务方在提供匹配查询结果的同时无法获知具体对应的个体信息。匿踪查询技术能够在数据不出库的同时支持多方联合建模计算，从而更安全的服务于金融信贷场景。例如，富数科技推出了集成匿踪查询技术能力的企业级安全计算平台 Avatar,目前已在银行、保险、消费金融等领开展应用。

(三)深度伪造应用安全防护

1.安全风险

2017 年，美国 Reddit 新闻网站上一位名为 deepfakes 的用户上传了经过技术篡改的色情视频，将视频中的演员人脸替换成电影明星的脸。由此“深度伪造（Deepfake）”技术引发人们关注。深度伪造技术不合理应用带来的安全风险主要体现在以下三个方面。

一是损害个体肖像权、名誉权与隐私权。随着深度伪造技术开源代码、APP 应用增多，深度伪造技术门槛不断降低。不法分子利用深度伪造技术制作虚假视频，可能被用于诬陷、诽谤和报复他人的手段，

侵犯个人肖像权、名誉权和隐私权。截至 2019 年 12 月，全网流传的深度伪造视频中，虚假色情内容占比高达 96%。

二是助长网络诈骗。传统诈骗手段在深度伪造技术加持下更加猖獗。例如，非法人员利用人工智能技术将自己的声音伪造成受害者信任人员的音色，通过语音聊天方式实施诈骗。2019 年德国某公司 CEO 因虚假伪造语音电话被骗取 220,000 欧元。

三是加剧网络谣言传播。深度伪造技术可被用于生产虚假新闻信息，成为网络谣言生产工具，助长网络谣言传播，冲击新闻媒体社会公信力。例如，2018 年特朗普宣布美国退出《巴黎气候协定》后不久，比利时某政党制作了一段特朗普呼吁比利时仿效美国的虚假讲话视频。尽管视频末尾注明了“这不是真的特朗普”，但该视频仍在比利时引起轩然大波。

2. 安全防护

(1) 技术检测

针对深度伪造滥用的防御方法，首要是使用技术手段对虚假伪造图像、视频、音频进行检测。现阶段学术界和工业界已经发布了多种针对深度伪造内容的检测方法。

早期深度伪造检测方法大多采用基于人为设计特征的图像取证技术。该类方法采用传统信号处理技术，使用图像频域和统计特征区分伪造内容和真实内容，例如通过局部噪音分析、图像质量评估、光照一致性等方法检测识别复制-移动、拼接、移除等图像篡改情况。深度伪造视频本质是一系列伪造图片的时序组合，因而可以将图像取

证技术方法应用于深度伪造视频检测。例如加利福尼亚大学提出了一种用于检测 Photoshop 修改人脸的方法，首先对输入真图自动生成假图，然后利用标注好的假图和真图训练分类网络，从而实现在假图中自动检测篡改内容位置的目标。然而，随着深度伪造技术的发展，合成内容效果越来越逼真，难以通过图像频域和统计特征实现对深度伪造合成内容的准确检测。

现阶段深度伪造检测方法主要采用基于深度神经网络的伪造特征自动提取检测技术。一是基于人脸检测和人脸特征提取的深度伪造检测方法。该类方法首先运用深度神经网络实现人脸检测和特征提取，然后利用篡改人脸特征和真实人脸特征的差异检测人脸是否遭受篡改。二是基于图像篡改痕迹的深度伪造检测方法。该类方法利用深度神经网络自主学习发现人脸篡改区域特征，并以此为依据实现伪造内容检测。例如，北京大学和微软亚洲研究院提出了 Face X-Ray 方法，通过使用深度神经网络学习发现图像融合边界，进而判断输入图像是否为篡改合成图像。由于基于深度神经网络的伪造特征自动提取检测技术只能准确检测发现在训练数据集中出现的类似伪造样例，对于新深度伪造方法生成的伪造内容检测效果往往不佳。

基于生理信号特征的深度伪造检测方法日益受到重视。该类方法通过比较深度伪造视频和真实视频之间在人物眨眼频率、心率、语速、声纹等生理信号特征差异检测深度伪造内容。例如，宾汉顿大学和 Intel 公司提出了 FakeCatcher 方法，通过测量脸部皮肤光电容积脉搏波信号实现伪造视频检测。佛罗里达大学提出了利用正常视频与异

常视频中人物心率分布不同检测深度伪造视频的方法。然而，由于心率、皮肤状态等生理特征易被外界因素干扰难以提取，进而导致检测准确率下降。

(2) 内容溯源

检测技术发展滞后于深度伪造自身技术的演进。不断更新迭代的深度伪造技术常使检测技术失效。因而需要从源头出发区分真实内容和伪造内容。我国《数据安全管理办法（征求意见稿）》第 24 条明确提出“网络运营者利用大数据、人工智能等技术自动合成新闻、博文、帖子、评论等信息，应以明显方式标明‘合成’字样”。美国《深度伪造责任法案》规定，利用深度合成技术合成虚假内容放置于网上传播的，制作者应当采用嵌入数字水印、文字、语音标识等方式披露合成信息。例如，2019 年 11 月 Twitter 公司指出将在推文下方附加“包含合成或篡改内容”的提示信息。Sensity 视觉威胁情报平台提供了深度伪造内容查询服务，从源头上追溯视频的真伪。

(3) 行业实践

美国谷歌、脸书、亚马逊等主流科技公司纷纷采取了技术措施防范深度伪造滥用。例如，脸书、亚马逊、微软联合学术界发起名为 Deepfake Detection Challenge (DFDC) 的挑战赛，悬赏深度伪造假视频的最佳检测方法。脸书对虚假视频进行标注，宣布了四种方法屏蔽虚假信息和仇恨言论，以减缓它们在社交网络上的传播速度。谷歌开源了包含 3000 个人工智能生成的虚假伪造视频数据集，助力打击深度伪造。Github 封杀了 DeepFake 和 DeepNude 等深度伪造应用的副

本。在国内，阿里巴巴安全图灵实验室宣布研发出针对换脸视频的深度伪造检测技术，这种方法标注简单，并能帮助神经网络更好的学习人脸特征，实现更好的检测效果。百度和瑞莱智慧推出了深度伪造检测服务平台，可向视频网站、网络论坛、新闻机构等提供人脸和人声伪造检测能力。

中国信息通信研究院安全研究所

地址：北京市海淀区花园北路 52 号

邮政编码：100191

联系电话：010-62305960

传真：010-62300264

网址：www.caict.ac.cn