计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300306

人工智能安全框架

景慧昀[1]　魏　薇[1]　周　川[2,3]　贺　欣[4]
1 中国信息通信研究院　北京 100083
2 中国科学院信息工程研究所　北京 100097
3 中国科学院大学网络空间安全学院　北京 100049
4 国家计算机网络应急技术处理协调中心　北京 102209
(jinghuiyun@caict.ac.cn)

摘　要　随着人工智能时代的到来,各行各业均开始结合自身业务需要部署人工智能系统,这全面加速了全球范围内人工智能规模化部署和应用进程。然而,人工智能基础设施、设计研发以及融合应用过程中面临的安全风险也随之而来。为了充分规避风险,世界各国纷纷采取制定人工智能伦理准则、完善法律法规和行业管理等方式来进行人工智能安全治理。在人工智能安全治理中,人工智能安全技术体系具有重要指导意义。具体而言,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,更是人工智能产业健康有序发展的重要保障。然而,在当前阶段,全球范围内人工智能安全框架普遍缺失,安全风险突出且分立,迫切需要对人工智能各生命周期存在的安全风险进行总结与归纳。为解决上述问题,文中提出了涵盖人工智能安全目标、人工智能安全分级能力、人工智能安全技术和管理体系的人工智能安全框架,期待为社会各界提升人工智能安全防护能力提供有益参考。

关键词:人工智能;安全框架
中图法分类号　TP183

Artificial Intelligence Security Framework

JING Hui-yun[1],WEI Wei[1],ZHOU Chuan[2,3] and HE Xin[4]
1 China Academy of Information and Communications Technology,Beijing 100083,China
2 Institute of Information Engineering,Chinese Academy of Sciences,Beijing 100097,China
3 School of Cyber Security,University of Chinese Academy of Sciences,Beijing 100049,China
4 National Computer Network Emergency Response Technical Team/Coordination Center of China,Beijing 102209,China

Abstract　With the advent of artificial intelligence,all walks of life begin to deploy artificial intelligence systems according to their own business needs,which accelerates the scale construction and widespread application of artificial intelligence worldwide in an all-around way. However,the security risks of artificial intelligence infrastructure,design and development,and integration applications also arise. To avoid risks,countries worldwide have formulated AI ethical norms and improved laws and regulations and industry management to carry out artificial intelligence safety governance. In the artificial intelligence security governance,the artificial intelligence security technology system has important guiding significance. Specifically,the artificial intelligence security technology system is an essential part of artificial intelligence security governance,critical support for implementing artificial intelligence ethical norms,meeting legal and regulatory requirements. However,there is a general lack of artificial intelligence security framework in the world at the current stage,and security risks are prominent and separated. Therefore,it is urgent to summarize and conclude the security risks existing in each life cycle of artificial intelligence. To solve the above problems,this paper proposes an AI security framework covering AI security goals,graded capabilities of AI security,and AI security technologies and management systems. It looks forward to providing valuable references for the community to improve artificial intelligence's safety and protection capabilities.

Keywords　Artificial intelligence,Security framework

到稿日期:2021-03-30　返修日期:2021-04-28
基金项目:国家 242 信息安全计划(2018Q39)
This work was supported by the National 242 Information Security Program(2018Q39).
通信作者:贺欣(hexin@cert.org.cn)

1 人工智能安全风险

结合人工智能系统设计运营等全流程，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，将有助于分析定位风险来源，研究和部署针对性的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 8 个阶段。基于 ISO 对人工智能系统全生命周期的划分，项目组描绘出了人工智能生命周期安全风险地图，如图 1 所示。

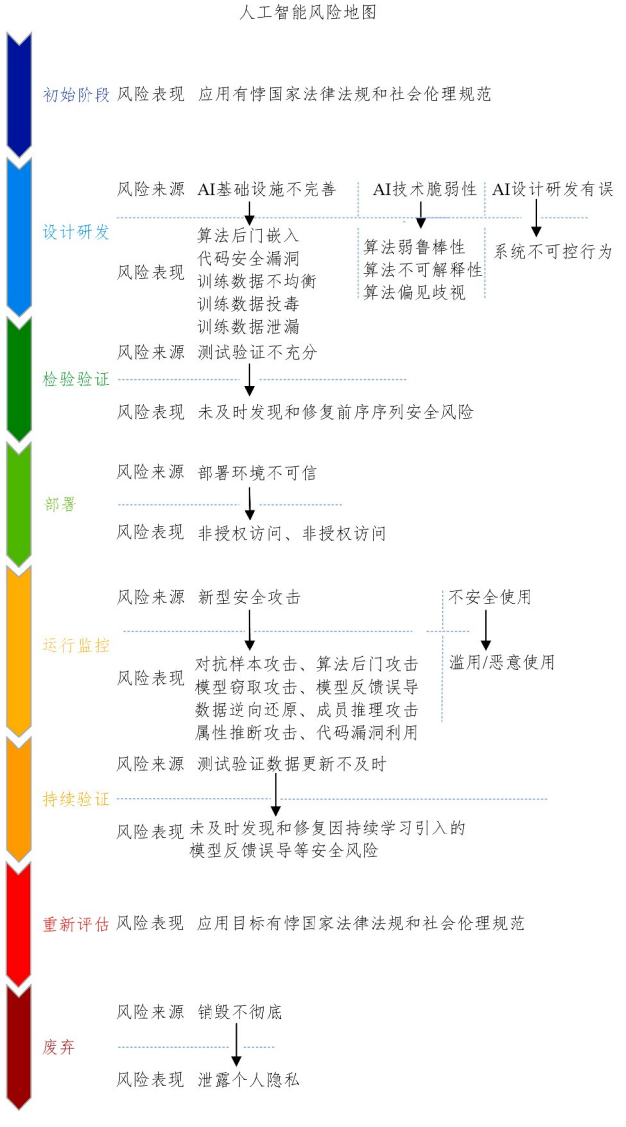


图 1 人工智能风险地图

Fig. 1 Security risk map of artificial intelligence

1.1 初始阶段安全风险

初始阶段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、风险管理设计等过程。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。

1.2 设计研发阶段安全风险

设计研发阶段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、编写软件代码、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[1-2]以及因设计研发有误等引发的安全风险。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[3]。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、个人隐私泄露等安全风险。

1.3 检验验证阶段安全风险

检验验证阶段主要负责检查人工智能系统是否按照预期要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、未能及时发现和修复前序阶段的安全风险[1-2]。

1.4 部署阶段安全风险

部署阶段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部署人工智能系统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[4-5]。

1.5 运行监控阶段安全风险

运行监控阶段主要包括针对人工智能系统的运行监控、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[6]、算法后门[7]、模型窃取[8]、模型反馈误导、数据逆向还原、成员推理[9-10]、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的风险。例如，在针对亚马逊和谷歌的在线人工智能分类任务进行黑盒对抗攻击时，研究者对窃取到的替代模型生成白盒对抗攻击样本，使用该样本使得亚马逊和谷歌的分类模型分别出现 96.19%和 88.94%的误判率[11]。

1.6 持续验证阶段安全风险

在持续验证阶段，针对持续学习的人工智能系统进行持续检验验证。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。

1.7 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重定义、需求重定义、风险管理重设计等过程，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。

1.8 废弃阶段安全风险

在废弃阶段，销毁使用目的不复存在或者有更好替代方案的人工智能系统，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[12]。

2 人工智能安全框架

2.1 设计思路

2.1.1 框架范围

本框架聚焦于人工智能内的安全，重点关注解决人工智能基础设施和设计研发面临的安全风险，以及因上述安全问题直接引发的人工智能应用发生行为决策失控的安全风险。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

社会安全风险，主要由国家法律法规和行业监管政策对使用者予以规制。

2.1.2 核心要素

基于人工智能安全框架应遵循实用性、前瞻性和整体性原则，从以下 3 个方面构建人工智能安全框架。

(1)明确人工智能安全目标是前提。本框架通过全面分析人工智能应用面临的安全风险，提出了人工智能安全目标，为人工智能安全防护工作的实施指明了方向。

(2)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[13]，提出了人工智能安全能力分级叠加演进模型。

(3)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 4 个维度，需要从 4 个不同的层面指导企业开展人工智能安全防护工作。

2.2 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 4 个维度，如图 2 所示。这 4 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，设定合理的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。

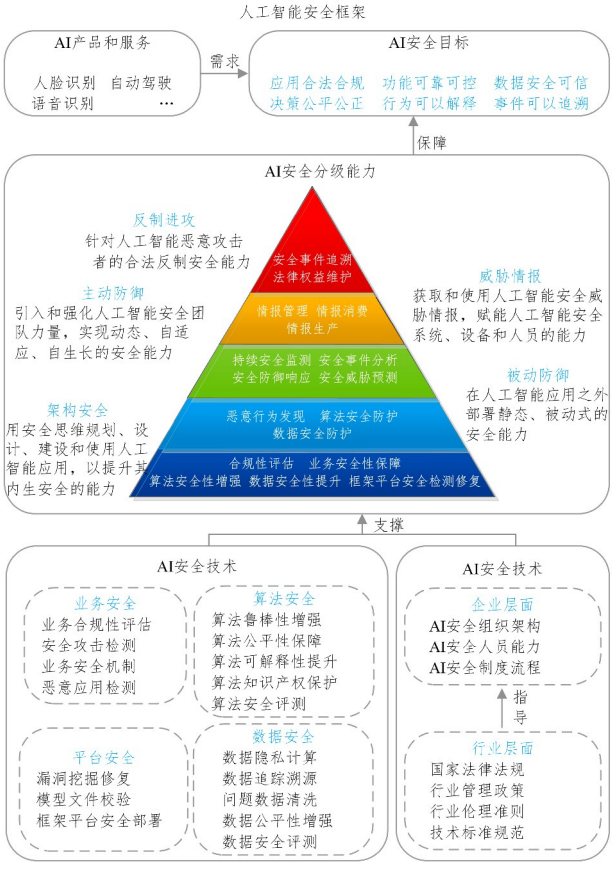


图 2 人工智能安全框架

Fig. 2 Artificial intelligence security framework

2.2.1 安全目标

通过系统地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 6 个方面提出安全需求和目标。

(1)应用合法合规。人工智能已在交通、医疗等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。

(2)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[1]、算法后门[2]等新型安全攻击方式，可通过实时修改输入数据诱使人工智能应用产生非预期的错误输出。因此，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的控制之下。

(3)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[14]、数据偏见、数据投毒[15]等诸多安全隐患。因此，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不被篡改，能够真实反映物理世界和人类社会的情况。

(4)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，从而损害社会的公平和正义。因此，应确保人工智能应用兼顾各类群体的特征信息，不会对特定的人或群体做出带有歧视和偏见的决策。

(5)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[16-18]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。因此，应确保人工智能应用以人类可以理解的方式对其行为和结果提供合理、准确的解释。

(6)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。因此，人工智能应用应根据业务场景量体裁衣，完善追溯体系，部署对安全事件产生的原因、发生环节、行为主体等进行有效追踪溯源的技术措施。

2.2.2 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 5 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。第二级安全能力是被动防御，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。第三级安全能力是主动防御，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。第四级安全能力是威胁情报，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。第五级安全能力是反制进攻，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5个方面。

1)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实现方式是否符合国家法律法规、行业监管政策以及伦理规范。

2)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，保障在遭遇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，保障危险情况下汽车控制权被及时交还给人类。

3)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，来增强算法鲁棒性、可解释性和公平性等。例如，可通过对抗训练[6]、模型正则化[19]等方式提升算法的鲁棒性。

4)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[20]、同态加密[21]、联邦学习[22]等技术提升数据的机密性。

5)框架平台安全检测修复。对来自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进行修复，以提前感知风险，降低安全事件发生的概率。例如，MindSpore[23]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3个方面。

1)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[6]、算法后门[7]、模型窃取[8]、深度伪造[24]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件或模型，利用正常样本和对抗样本在特征空间中的差异检测来对抗样本攻击。

2)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对抗样本[6]、模型窃取[8]等算法的安全攻击。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，从而达到防御对抗样本攻击的目的。

3)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地抵御训练数据投毒、数据逆向还原、成员推理[9-10]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的过程，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性，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4个方面。

1)持续安全监测。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，给出应用当前的安全风险级别，并对应用运行异常及时告警。

2)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，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。

3)安全防御响应。在安全事件发生时，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。

4)安全威胁预测。运用人工智能、大数据分析等技术，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进一步提升和扩展主动防御效能。威胁情报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，主要包括以下3个方面。

1)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。

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

3)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能力，主要包括以下两个方面。

1)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以支撑后续的法律权益维护。

2)法律权益维护。出于自卫的目的，运用法律手段对攻击者采取反击行为。

2.2.3 安全技术

人工智能业务、算法、训练数据和机器学习框架平台是构建人工智能应用的4个核心组件，也是人工智能安全的重点防护对象。因此，本框架针对业务、算法、数据和平台提出安全防护技术手段。

2.2.4 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，体现在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并遵守人工智能安全管理规则的基础上，不断完善人工智能安全管理组织机构配置、制度流程和提升人员能力。本文聚焦于企业层面安全管理体系，主要包括以下3个方面。

(1)建立安全管理组织架构。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，明确岗位职责与人员分配。

(2)增强安全管理人员能力。增强和扩大人工智能安全专业人才队伍，提升从业人员的专业技能。

(3)制定安全管理流程制度。制定人工智能安全总纲、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。

3 人工智能安全技术

近期，人工智能安全领域受到了全球的重点关注，已在算法鲁棒性增强、可解释性提升以及算法安全攻击检测和防御技术等方面取得了局部突破，可支撑实现在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。人工智能安全技术图谱如图 3 所示。

图 3 人工智能安全技术图谱

Fig. 3 Artificial intelligence security technology atlas

3.1 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，主要包括以下 4 个方面。

(1)业务合规性评估。基于国家法律法规、行业监管政策以及人工智能伦理规范等针对人工智能技术和应用提出的安全要求，由人工智能安全专家对企业研发的面向具体业务场景的人工智能系统的应用目标和方式进行合规性评估。

(2)安全攻击检测。通过对人工智能应用的输入数据、访问行为、输出结果等方面进行综合分析，持续检测人工智能应用可能遭受的安全攻击，及时发现安全事件，支撑后续的安全事件响应处理。这一部分包含对抗样本检测[25-27]、模型后门检测[28-29]、成员推理检测、属性推断检测、数据逆向还原检测与模型窃取检测[8,30]。

目前，对于成员推理、属性推断、数据逆向还原等安全攻击，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的方法研究尚未获得广泛关注。

(3)恶意应用检测。综合运用多媒体分析识别、数字取证等技术对虚假伪造信息生成、验证码绕过等人工智能恶意应用行为进行分析识别，例如针对深度合成图像 Deepfake 的研究获得了广泛的关注[31-32]。

(4)业务安全机制。综合运用身份验证、访问次数限制、模块隔离、多级安全机制、系统监测等传统网络安全技术措施，实现业务访问控制、安全隔离、安全熔断和安全监控等安全机制。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。

3.2 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，主要包括以下 5 个方面。

(1)算法鲁棒性增强。人工智能算法模型的鲁棒性主要面临两个方面的挑战。一是正常环境扰动对算法模型鲁棒性的影响，例如拍摄光线亮度、角度、距离等都对机器视觉算法的鲁棒性造成了挑战。二是恶意攻击者对训练数据和运行输入数据进行的特定扰动对算法模型鲁棒性产生影响，例如训练数据投毒、对抗样本攻击等均可导致人工智能算法模型产生错误输出。综合运用数据增强、鲁棒特征学习、模型随机化、模型正则化等方法，能够有效提升人工智能算法模型的鲁棒性。例如，Madry 等[33]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能力，进而提升模型鲁棒性。

(2)算法公平性保障。在算法层面实施的保障公平性的技术措施主要有两类。一是改进人工智能算法自身，例如通过对算法附加公平性约束条件、减少算法对敏感属性的依赖、偏见后处理等方法增强算法自身的公平性。二是开展人工智能算法公平性审计，设计算法公平性审计数据集和审计指标，客观评价算法的公平性。例如，Bose 等[34]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。

(3)算法可解释性提升。提升人工智能算法可解释性的方法主要有两类。一是算法模型自身可解释，对于已经训练好的算法，模型无需额外的信息就可以理解模型的决策过程或决策依据，如朴素贝叶斯、线性回归、决策树、基于规则的模型等。二是算法模型事后解释技术[35]，即利用解释方法或构建解释模型对已经训练好的算法模型的工作机制、决策行为和决策依据进行解释。例如，Liu 等[36]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，以降低模型的复杂度，进而更好地对算法进行解释。Yang 等[37]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。

(4)算法知识产权保护。在算法模型训练时将水印嵌入模型文件的算法水印技术,实现了对算法模型窃取行为的追踪溯源。例如,Zhang 等[38]提出了首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/输出在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。

(5)算法安全评测。综合运用公平性、鲁棒性、可解释性等方面的安全评价指标和评测技术工具,全面分析人工智能算法面临的安全缺陷和问题。

3.3 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,主要包括以下 5 个方面。

(1)数据隐私计算。其指在不泄露数据信息的前提下实现数据分析计算的一类技术,可提升人工智能应用数据的机密性,有效防御训练数据窃取、成员推理攻击、数据逆向还原等攻击。依据技术原理的不同,数据隐私计算主要分为密码学、可信执行环境、联邦学习[22]3 个方向。其中,密码学方向以安全多方计算、同态加密[21]、零知识证明、差分隐私[20]等技术为代表。可信执行环境指构建硬件安全区域,数据仅在该安全区域内进行分析计算,可实现隐私数据的安全存储隔离、安全传输、安全计算和删除。联邦学习是一种分布式机器学习框架,能在原始数据保留在各计算参与方本地的情况下,实现机器学习协同建模的目的。

(2)数据追踪溯源。其指追踪数据来源并重现数据历史处理状态的一类技术,可增强人工智能应用的训练数据的完整性和真实性,为及时发现遭投毒篡改的训练数据提供可靠机制。目前,数据安全标签、区块链是两类常用的数据追踪溯源方法。数据安全标签是一种简单、有效且应用广泛的数据溯源方法,通过将记录数据溯源信息的标签和数据一起传播,可达到追溯数据历史状态、发现数据篡改行为的目的,是保障数据溯源信息不被破坏篡改和实现数据溯源的关键。利用区块链技术的数据不可篡改、可溯源的特点,通过区块链保存用于记录训练数据集来源以及演变过程的溯源信息,可更好地达到验证训练数据真实性和完整性的目的。

(3)问题数据清洗。其指对不完整数据、冗余数据、冲突数据和错误数据进行检测和纠正的一类技术,可提升人工智能应用的数据可用性,为防御训练数据投毒、对抗样本攻击等提供了有效方式。数据清洗处理主要包括问题数据检测删除和问题数据重构修复两大类。问题数据检测删除方法利用问题数据和正常数据的分布不同,来达到检测发现和丢弃问题数据的目的。问题数据重构修复方法通过比对异常数据与正常数据的差异来对异常数据进行重构修复,在正常数据匮乏的情况下可利用修复重构后的数据进行模型训练和预测。例如 Tran 等[39]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,并过滤掉这些投毒数据。

(4)数据公平性增强。其指消除训练数据集中存在的种族偏差、样本选取偏差等可能导致算法模型不公平决策的一类技术,可提升人工智能应用的公平性。数据公平性增强主要包括修正训练数据比例、训练数据转换等方法。修正训练数据比例通过训练数据增广、组合不同来源的数据等方式均衡具有不同敏感属性值和类标签的训练数据所占的比例。例如 Feldman 等[40]对训练数据的每个属性进行修改,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[41]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。基于国家和行业数据安全法律法规及技术标准,通过专家审查以及模拟数据窃取攻击等方式,对人工智能应用的数据合规性以及数据安全性进行评测。

3.4 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,主要包括以下 3 个方面。

(1)漏洞挖掘修复。在使用人工智能开源框架平台前,通过静态代码审计、动态模糊测试等漏洞挖掘分析技术,及时发现人工智能开源框架平台的代码安全漏洞和隐患。人工智能开源框架平台社区设立安全问题反馈机制,及时组织研发人员接受并处理修复内外部安全研究人员发现提交的安全问题。

(2)模型文件校验。为防止人工智能开源框架平台加载恶意的人工智能模型文件,通过检查模型文件的格式、大小、参数范围、网络结构、模型来源等信息,及时发现可疑的被投毒或者被篡改的模型文件。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。

(3)框架平台安全部署。通过将人工智能开源框架平台部署在可信环境或者权限受限的沙箱环境中,能够确保人工智能开源框架平台不会泄露人工智能应用的数据和算法隐私,且无法通过承载框架的操作系统攻击人工智能开源框架平台。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,则可能导致任意代码执行的严重后果。

结束语 本文总结了人工智能的基础设施、设计和研发以及集成应用所面临的日益严峻的安全风险,并相应绘制出了一幅人工智能安全风险地图,该地图详细刻画出了人工智能全生命周期中各个阶段面临的安全风险的来源和表现。然而,由于全球范围内缺乏人工智能安全框架,无法支撑人工智能安全治理,不能对人工智能产业的健康有序发展提供有力支撑。针对当前人工智能面临的突出安全风险,我们提出了涵盖人工智能安全目标、人工智能安全分级能力、人工智能安全技术和人工智能管理系统的人工智能安全框架。希望为社会各界提高人工智能的安全防护能力提供有益的参考。

参 考 文 献

[1] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv: 1712.05526, 2017.

[2] PEI K, CAO Y, YANG J, et al. Deepxplore: Automated white-box testing of deep learning systems[C]// Proceedings of the 26th Symposium on Operating Systems Principles. 2017: 1-18.

[3] 360政企安全. AI框架安全依旧堪忧：360 AI安全研究院披露Tensorflow 24个漏洞[EB/OL]. http://www.anquanke.com/post/id/218839.

[4] ZHANG T, HE Z, LEE R B. Privacy-preserving machine learning through data obfuscation[J]. arXiv:1807.01860, 2018.

[5] XU K, CAO T, SHAH S, et al. Cleaning the null space: A privacy mechanism for predictors[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. 2017: 2789-2795.

[6] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv:1312.6199, 2013.

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

[8] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction apis[C]// 25th {USENIX} Security Symposium ({USENIX} Security 16). 2016: 601-618.

[9] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models[C]// 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017: 3-18.

[10] FREDRIKSON M, LANTZ E, JHA S, et al. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing[C]// 23rd {USENIX} Security Symposium ({USENIX} Security 14). 2014: 17-32.

[11] REN K, MENG Q R, YAN S K, et al. Survey of artificial intelligence data security and privacy protection[J]. Chinese Journal of Network and Information Security, 2021, 7(1): 1-10.

[12] CHEN D, ZHAO H. Data security and privacy protection issues in cloud computing[C]// 2012 International Conference on Computer Science and Electronics Engineering. IEEE, 2012, 1: 647-651.

[13] ROBERT M L. The Sliding Scale of Cyber Security[EB/OL]. http://www.sans.org/reading-room/whitepapers/analyst/mem-bership/36240, 2015-8.

[14] YEOM S, GIACOMELLI I, FREDRIKSON M, et al. Privacy risk in machine learning: Analyzing the connection to overfitting[C]// 2018 IEEE 31st Computer Security Foundations Symposium (CSF). IEEE, 2018: 268-282.

[15] JAGIELSKI M, OPREA A, BIGGIO B, et al. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning[C]// 2018 IEEE Symposium on Security and Privacy (SP). IEEE, 2018: 19-35.

[16] GILPIN L H, BAU D, YUAN B Z, et al. Explaining explanations: An overview of interpretability of machine learning[C]// 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2018: 80-89.

[17] HUA Y, ZHANG D, GE S. Research progress in the interpretability of deep learning models[J]. Journal of Cyber Security, 2020, 5(3): 1-12.

[18] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]// European Conference on Computer Vision. Springer, Cham, 2014: 818-833.

[19] TALVITIE E. Model Regularization for Stable Sample Rollouts[C]// UAI. 2014: 780-789.

[20] DWORK C. Differential privacy: A survey of results[C]// International Conference on Theory and Applications of Models of Computation. Berlin, Heidelberg: Springer, 2008: 1-19.

[21] GENTRY C. A fully homomorphic encryption scheme[M]. Stanford: Stanford University, 2009.

[22] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2): 1-19.

[23] MindSpore[OL]. http://www.mindspore.cn/security.

[24] Porn Producers Offer to Help Hollywood Take Down Deepfake Videos, Janko, Roettgers[OL]. http://variety.com/2018/digital/news/deepfakes-porn-adult-industry-1202705749/, Aug. 2019.

[25] MENG D, CHEN H. Magnet: a two-pronged defense against adversarial examples[C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 135-147.

[26] MA X, LI B, WANG Y, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[J]. arXiv: 1801.02613, 2018.

[27] GU S, YI P, ZHU T, et al. Detecting Adversarial Examples in Deep Neural Networks using Normalizing Filters[C]// 11th International Conference on Agents and Artificial Intelligence. 2019.

[28] CHEN B, CARVALHO W, BARACALDO N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[J]. arXiv: 1811.03728, 2018.

[29] WANG B, YAO Y, SHAN S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C]// 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019: 707-723.

[30] OH S J, SCHIELE B, FRITZ M. Towards reverse-engineering black-box neural networks[M]// Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, Cham, 2019: 121-144.

[31] CHEN Z, XIE L, PANG S, et al. MagDR: Mask-guided Detection and Reconstruction for Defending Deepfakes[J]. arXiv: 2103.14211, 2021.

[32] ZHAO H, ZHOU W, CHEN D, et al. Multi-attentional deepfake detection[J]. arXiv: 2103.02406, 2021.

[33] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv: 1706.06083, 2017.

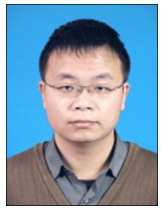
[34] BOSE A, HAMILTON W. Compositional fairness constraints for graph embeddings[C]// International Conference on Machine Learning. PMLR, 2019: 715-724.

[35] CHAKRABORTY S, TOMSETT R, RAGHAVENDRA R, et al. Interpretability of deep learning models: a survey of results[C]// 2017 IEEE smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI). IEEE, 2017: 1-6.

[36] LIU X, WANG X, MATWIN S. Improving the interpretability of deep neural networks with knowledge distillation[C]// 2018 IEEE International Conference on Data Mining Workshops (IC-

DMW). IEEE, 2018: 905-912.
[37] YANG C, RANGARAJAN A, RANKA S. Global model interpretation via recursive partitioning[C]// 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2018: 1563-1570.
[38] ZHANG J, CHEN D, LIAO J, et al. Model watermarking for image processing networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(7): 12805-12812.
[39] TRAN B, LI J, MADRY A. Spectral signatures in backdoor attacks[J]. arXiv: 1811.00636, 2018.
[40] 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. 2015: 259-268.
[41] BELLAMY R K E, DEY K, HIND M, et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias[J]. arXiv: 1810.01943, 2018.

JING Hui-yun, born in 1987, Ph.D, senior engineer. Her main research interests include artificial intelligence security and data security.



HE Xin, born in 1982, Ph.D, senior engineer. His main research interests include network information security and so on.