人工智能应用系统
内生安全蓝皮书
(2024年)

中国通信学会内生安全专业技术委员会

2024年11月

版权声明

专家组和撰写组名单 顾问

(以姓氏笔划为序):

邬江兴

专家组：

组长：

张　帆　国家数字交换系统工程技术研究中心教授

副组长：

黄瑞阳　国家数字交换系统工程技术研究中心教授

成员(以姓氏笔划为序):

姓名	单位	职务
陈　曦	国家数字交换系统工程技术研究中心	助理研究员
郭　威	国家数字交换系统工程技术研究中心	副教授
邹　宏	复旦大学大数据研究院	副院长
尚玉婷	复旦大学大数据研究院	助理研究员
李建鹏	郑州大学	助理研究员
李邵梅	国家数字交换系统工程技术研究中心	副研究员
高彦钊	国家数字交换系统工程技术研究中心	副研究员

撰写组(按单位排名)

单位	姓名
紫金山实验室	杜加玉
紫金山实验室	黄 炜

紫金山实验室	苗馨远
紫金山实验室	曹 毅
紫金山实验室	董方旭
紫金山实验室	杨秋龙
紫金山实验室	朱 进
紫金山实验室	黄 潇
紫金山实验室	陈 鑫
紫金山实验室	周志中
紫金山实验室	彭自文
紫金山实验室	乔明起

序 言

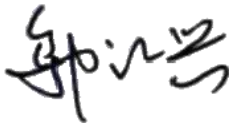
当前人工智能(Artificial Intelligence,AI)技术已经成为推动社 会经济发展的新引擎。然而，正如哲学家黑格尔指出的那样：“一切 事物都是自在性矛盾的”,任何伟大的技术发明在创造前所未有的机 遇同时，也一定会衍生或伴随前所未有的挑战。当前数据投毒、算法 偏见、模型脆弱性和运行环境内生安全威胁等问题频发，不仅影响到 AI应用系统可用性和可靠性，还能会触及隐私保护、网络安全、社 会稳定乃至生命财产安全和认知混乱等可信性方面的技术伦理议题。 因此，构建一套能够抵御内外部威胁、确保AI健康可持续发展的安 全体系，已成为智能时代不可或缺的刚需，必须防范重蹈“先发展后 治理”陈旧理念和模式覆辙。

哥德尔不完备性定理指出，即使是最精密细致设计的系统，也无 法证明不存在不确定性风险。AI 应用系统涵盖了从数据治理、模型 训练、部署应用，到运行监控及反馈优化等环节，每一个步骤都有可 能引入不确定性因素，主要表现在：一是AI 应用系统运行环境仍基 于冯·诺依曼的存储程序控制机理或架构，不可避免的存在“娘胎里 带来基因缺陷”;另一方面，AI 模型算法特有的“黑盒”效应具有 不可解释性、不可判识性和不可推论性，使其成为悬在任何实用化的 AI应用系统之上的“达摩克利斯之剑”。

现有AI 安全防御方法主要依赖“亡羊补牢，吃一堑长一智”的 经典反脆弱性方法论，这种策略尽管能够在一定程度上防范已知威胁，

但任何附加的安全算法或措施都无法杜绝自身存在的内生安全问题， 因而不可能从根本上规避“反复踩坑的恶性循环”,有效抵御广义不 确定性安全威胁。网络内生安全理论的提出，建立在必要多样性定律 和相对正确公理基础上的“构造决定安全”方法，能够在AI应用系 统设计之初就导入动态异构冗余构造和策略裁决功能，使其具备应对 任何差模性质不确定性影响的能力，可为提升AI应用系统网络韧性 或广义功能安全提供全新的视角和解决方案。

智能时代的技术实践需要正确的安全理论指引，在设计人工智能 应用系统时，不仅要充分考虑网络韧性架构设计，还要为适应可能出 现的各种未知安全风险留有足够余地。这既是人工智能技术发展的必 然要求，也是防范社会性系统风险的必要条件。本蓝皮书是一篇集科 学性、实用性和前瞻性于一体的报告，不仅能为学术界提供宝贵的理 论参考，而且可为产业界提供一套行之有效的安全指南，更可以为一 体化的设计“高可信、高可靠、高可用”AI 应用系统提供一个具有 普适性意义的解决方案，帮助我们在未来的发展道路上走得更加稳健，让“泛在化AI技术向善”。



前 言

2024年1月，人工智能(AI) 新锐巨头OpenAI 公司的创始人兼 首席执行官萨姆·奥尔特曼(Sam Altman)在世界达沃斯论坛上强调， AI带来的技术革命不同于以往，而是成为了一种“不可思议的提高 生产力的工具”。作为最具颠覆性的新兴技术之一，AI 发展在为人类 社会经济发展带来巨大红利的同时，也引发了一系列现实危害和风险 挑战，主要包括数据窃取、隐私泄露、算法歧视、对抗攻击等安全威 胁，它们不仅威胁着个人信息的保护，影响着企业的信誉和运营安全， 甚至对国家安全也构成了潜在的隐患，这些问题需要我们持续关注并 采取有效措施加以防范和解决。

当前人工智能应用系统的安全威胁是由于其网络软硬件运行环 境和模型算法两个层面存在的“内生安全问题”所致。然而现有的人 工智能应用系统安全防护模式和技术路线很少能跳出“尽力而为、问 题归零”的惯性思维，传统的对抗训练、挖漏洞、打补丁、封门补漏、 查毒杀马乃至设蜜罐、布沙箱等层层叠叠的附加式防护措施，在引入 安全功能的同时不可避免地会引入新的内生安全隐患。

为创造性破解人工智能应用系统内生安全难题，本蓝皮书提出一 种内生安全赋能人工智能应用系统构建的方法，利用内生安全机理中 内在的构造效应，从体制机理上管控或规避人工智能应用系统面临的 破坏及威胁，进而有效提升人工智能应用系统在面对复杂多变安全环 境时的应对能力，为AI 技术的健康发展提供坚实保障。

目 录

一、研究概述

(一)人工智能应用系统

当前，全球人工智能技术取得了突飞猛进的发展，对经济社会发 展和人类文明进步的深远影响日益显现，给世界带来了前所未有的巨 大机遇。特别是以OpenAI 公司先后发布的ChatGPT、GPT-4o、GPT-o1 等为代表的大模型迅速崛起，更是将人工智能技术的发展推向了一个 新的高度，预示着一个更加智能、更加全面的智能时代的到来。当前 人工智能正加速与传统行业深度融合，给人机交互、客户服务、教育 辅导等多个领域带来了重大变革，成为推动社会经济转型升级和高质 量发展的重要力量。

图1显示了人工智能产业链框架，主要分为上游的基础层、中游 的技术层和下游的应用层。其中基础层是人工智能产业的基础，主要 指硬件设备和数据服务，为人工智能提供算力支持和数据基础，代表 性企业有英伟达、百度、地平线机器人等；技术层是人工智能产业的 核心，主要指通用技术、算法模型和开发平台，为人工智能产业提供技术支持，包含计算机视觉、自然语言处理、类脑算法、音频处理技 术、人机交互五类，代表性企业有OpenAI、谷歌、微软、英伟达、 百度、华为、迅飞、旷视科技、智谱华章等；下游的应用产品和应用 场景包括所有AI 技术与传统应用结合形成的产业种类，主要有语言 终端、智能汽车、机器人、视觉产品、智慧教育、智慧医疗、智能制 造等。

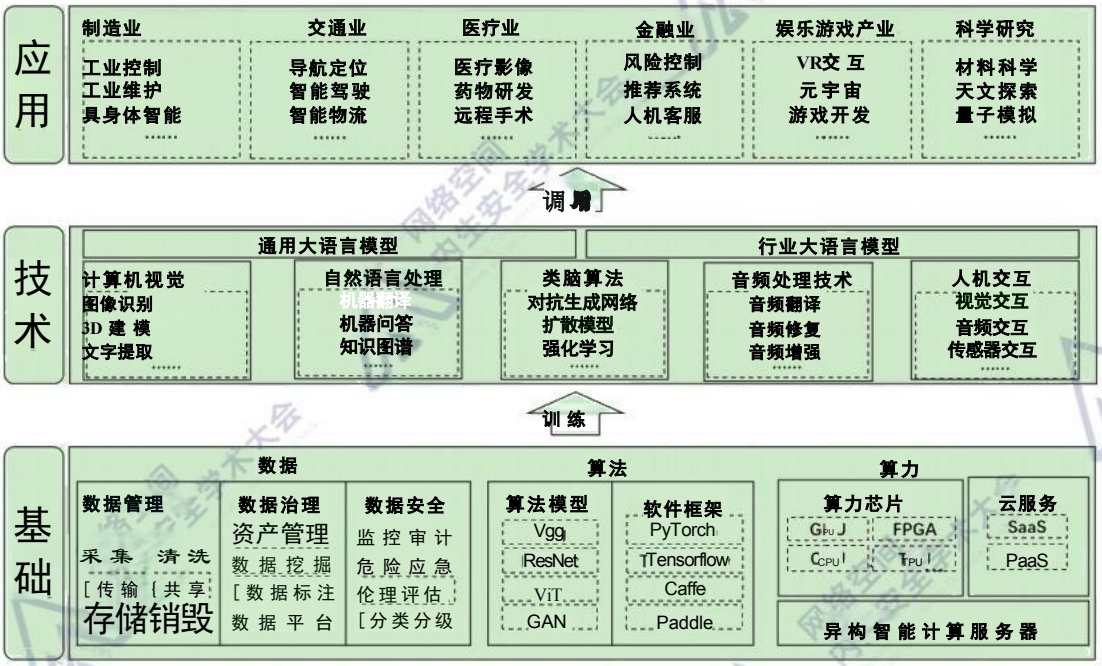


图1人工智能产业链

1.人工智能应用发展趋势分析

在科技飞速发展的今天，人工智能这股新浪潮正以前所未有的速 度冲击着各个领域，改写着人类的生活方式。从ChatGPT 为代表的 大语言单模态模型到AI文生图等多模态生成模型，从煤矿到铁路，

从金融到教育，AI 技术的影子无处不在。人工智能作为当今世界科 技革命和产业变革的关键领域，其产业应用的发展趋势呈现出多元化、 深度渗透和广泛应用的特点。

(1)技术创新持续升级

随着计算能力的不断提高和逻辑算法的不断优化，人工智能技术 将更加成熟。深度学习、强化学习、神经网络、语言大模型、多模态 预训练大模型等核心技术的不断进步，提升了语音识别、图像处理、 自然语言理解等领域的准确性和效率。当前多模态预训练大模型正成 为AI 大模型的主流形态，堪称当今人工智能产业的“标配”。

(2)行业应用加速扩展

人工智能正应用于众多行业和领域，成为推动传统产业升级和新 兴产业成长的重要力量。智能制造、智能医疗、智能交通、智慧城市 等已经成为人工智能应用的热点领域。特别是生成式人工智能技术的 飞跃式发展，其在广告营销、游戏创作、艺术设计等创造性工作场景 与行业中，正得到更为广泛的应用。一方面，创意属于稀缺资源，人 工智能生成内容(Artificial Intelligence Generated Content,AIGC)的 创造性对激发灵感、辅助创作、验证创意等大有裨益，另一方面，互 联网大规模普及使得“一切皆可在线”,数字内容消费需求持续旺盛，AIGC能更低成本、更高效率地生产内容，其经济性愈发凸显。

(3)智能化服务成为常态

智能客服、智能助手等智能化服务将更加普及，成为人们日常生 活和工作中不可或缺的一部分。同时，基于人工智能的大数据分析能 力，企业将能够提供更加精准和个性化的服务。由于ChatGPT 在文 本对话领域表现出的和人类行为的相似性，其被认为是人类通往通用 人工智能(Artificial General Intelligence,AGI) 道路上的里程碑式产 品。2023年7月，AI 科学家李飞飞团队公布了利用大型语言模型 (Large Language Models,LLMs)和视觉语言模型(Vision-Language Models,VLMs) 驱动的机器人项目 VoxPoser, 人类可以用自然语言 给机器人下达指令，机器人直接能够理解指令语义并做出相应动作， 而无需额外数据和训练。这些技术将推动智能化服务产业进入拓宽加 深期。

(4)伦理安全重视程度提高

随着人工智能技术的深入应用，深度神经网络大模型的预训练和 大规模人机交互中的强化学习技术正在推动人工智能向认知发展导

向的“自我进化”。然而，如何确保这种“自我特征”对人类社会有益 无害，以及如何保障数据安全和隐私，成为当前人工智能产业发展面 临的重大挑战。2023年11月，包括英国、美国、欧盟、澳大利亚和 中国等在内的28个国家和地区共同签署了《布莱切利宣言》,旨在 关注未来强大人工智能模型可能对人类生存构成的威胁，以及当前人 工智能增强有害或偏见信息的风险。这一宣言有力印证了当前全球对 人工智能安全伦理问题的重视。

(5)算力资源需求持续增长

随着人工智能应用的不断普及，计算能力的需求将持续增长。云 计算、边缘计算等新型计算模式将得到更广泛应用，以满足人工智能 大规模的算力需求。新硬件、新架构竞相涌现，现有芯片、操作系统、 应用软件等都可能被推翻重来，预计未来将实现“万物皆数据”“无 数不计算”“无算不智能”的愿景，届时智能算力将无处不在，呈现 “多元异构、软硬件协同、绿色集约、云边端一体化”四大特征。

(6)应用安全问题日益突出

随着人工智能技术的广泛应用，相关的安全隐患和挑战愈发严重。 对抗攻击、后门攻击、投毒攻击等多种安全问题揭示了人工智能技术 的安全缺陷，对个人和社会具备潜在的风险和隐患。各国和组织正加 紧制定监管措施，建立伦理框架和安全标准，以确保人工智能应用过 程中的透明性和可控性。但是，因为目前人工智能具有不可解释性、 不可判别性、不可推论性等特点，使得其安全问题难以从根本上解决， 因此亟需一个新的解决思路和框架。

2.人工智能应用系统分析

基于人工智能算法模型构建的应用系统体现在众多行业中，包括

城市公共安全管理、金融服务、医疗保健、零售和电子商务、生产制 造、交通物流、教学教育、娱乐旅游、食品科技、制造业、房地产业、 零售业等等。利用人工智能系统可以明显地提升工作效率、提高生产 力、优化工作流程并提高产品质量。以下是主要行业领域的人工智能 应用系统实例。

(1)城市公共安全领域

人工智能在提高城市公共安全保障能力方面优势明显，随着人工 智能发展所依赖数据的规模不断增大、算力的不断提高以及算法的不 断优化与突破，应用人工智能技术的成本将大幅降低，人工智能应用 系统正在城市公共安全领域发挥重要作用。具体包括：①视频智能监 控分析系统：人工智能可以应用于视频监控系统，通过图像识别、目 标检测和行为分析等技术，实时监测和分析监控视频中的异常行为，如盗窃、暴力等，发出警报并提供相关信息。例如在张学友2018年 的巡回演唱会上，民警便是利用演唱会的智慧安保人像识别功能，先 后追踪到80多名犯罪分子，顺利将他们抓捕归案。②社交媒体监测 系统：人工智能可以应用于社交媒体监测系统，通过自然语言处理和 情感分析等技术，实时监测和分析社交媒体上的言论和事件，发现恶 意言论、虚假信息等潜在的公共安全风险，并及时采取相应措施。③ 城市大数据分析与预警系统：人工智能可以应用于城市公共安全的大 数据分析与预警系统，通过整合和分析多种数据源，如监控数据、社 交媒体数据、气象数据等，发现异常模式和趋势，提前预警潜在的安 全风险，为决策者做出及时的应对措施提供支持。

(2)金融服务领域

人工智能技术在量化交易、风险评估、信贷评分等领域中广泛应

用。主要应用系统有：①金融风险评估系统：该系统利用人工智能和 大数据分析技术来评估金融领域风险，它通过分析大量的金融数据和 市场信息，利用人工智能算法，帮助金融机构和投资者识别、评估和 管理各种金融风险，以支持金融决策和风险管理。②欺诈检测系统： 通过分析大量的金融数据和交易信息，帮助金融机构和支付服务提供 商发现潜在的欺诈行为，准确识别信用卡欺诈、网络诈骗、洗钱等不 同类型的欺诈模式，并及时采取措施减少欺诈风险。③智能投资系统： 该系统能够根据投资者的风险偏好、投资目标和市场情况，利用人工 智能技术为投资者提供个性化投资建议和服务，系统会根据投资者的 目标和风险偏好，选择适合的资产类别、投资品种和配置比例，帮助 投资者更好地理解市场趋势和风险，提供更准确的投资建议，同时实 现交易的自动化执行。

(3)医疗保健领域

人工智能技术在影像诊断、疾病预测、智能医疗设备等领域中广 泛应用。主要应用系统有：①医学影像分析系统：利用深度学习和计 算机视觉技术，对医学影像进行自动分析和诊断，如肿瘤检测、疾病 分类等。②个性化医疗推荐系统：根据患者的病历数据和基因组信息， 人工智能技术可以提供更精确的诊断和治疗方案，并在早期发现疾病， 可以辅助医生进行疾病诊断，提供更准确的诊断结果和治疗建议，为 患者提供个性化的治疗方案和药物推荐。③医疗机器人系统：结合机 器人和人工智能技术，用于手术辅助、康复治疗和护理服务等。人工 智能可以模拟医生的思维和诊断推理，通过大量学习医学影像，可以 帮助医生进行病灶区域定位，给出患者可靠的诊断和治疗方案，减少 漏诊误诊的问题；同时像骨科、泌尿外科、妇产科、神经外科等多个

医学领域中，手术机器人能够极大地提高手术的精准率而被广泛使用。 ④智能医疗大数据分析系统：利用人工智能和大数据分析技术来处理 和分析医疗领域大规模数据的系统，它通过对医疗数据的深度分析和 挖掘，为医疗服务提供者、患者和医疗决策者提供精准、高效的决策 支持，如传染病监测与防控策略制定、药物研发支持等。

(4)零售和电子商务领域

人工智能技术可以帮助零售商和电商平台提供更好的产品和服 务，提高销售效率和用户体验。以下是一些常见的人工智能应用系统： ①智能推荐系统：推荐系统利用机器学习和数据分析技术，通过分析 用户的购买记录、点击行为和社交媒体数据，构建用户的兴趣和需求 模型，根据用户的历史行为和偏好，为他们推荐个性化的产品和服务， 提高销售转化率和用户满意度。②机器人客服：聊天机器人利用自然 语言处理和机器学习技术，与用户进行自动化的对话交互，如在电商 领域，机器人客服可以回答用户的问题、提供产品建议、处理订单和 退款等事务性任务，提供即时的客户服务和支持，提高用户体验和客 户满意度。③智能供应链管理系统：利用数据分析和预测算法，优化 零售和电商平台的供应链管理，如通过分析销售数据、库存情况和交 通运输信息，供应链优化系统可以预测需求、优化库存管理、提高配 送效率，减少成本和提供更快的交付服务。

(5)生产制造领域

人工智能技术在智能装配线、智能机器人等领域中广泛应用。主 要应用系统包括：①智能生产调度系统：通过实时监测和分析生产数 据，可以帮助发现生产流程中的瓶颈，提出改进建议，优化生产计划 和资源分配，提高生产效率和质量。②预测维护系统：利用机器学习

和传感器数据，预测设备故障和维护需求，从而提前进行维护，减少 停机时间和维修成本。③自动化智能机器人系统：结合机器人和人工 智能技术，使得生产制造领域的机器人能够在更复杂的环境中工作， 执行更精细的任务，如组装、包装、搬运等，实现自动化的生产和装 配，提高生产线的灵活性和效率。

(6)交通和物流领域

人工智能技术在交通调度、智能驾驶等领域广泛应用。主要应用 系统包括：①智能交通管理系统：利用实时交通数据和优化算法，实 现智能信号控制和交通流量优化，减少拥堵和提高道路利用率，并能 够预测未来一段时期内的交通模式，提前规划交通流量。②自动驾驶 技术：以雷达、激光雷达等传感器等基础，结合计算机视觉和机器学习，实现自动驾驶车辆的感知、决策和控制，使自动驾驶汽车可以在 没有人类干预的情况下导航和驾驶，提高交通安全和效率。③物流智 能调度系统：通过数据分析和路径规划算法，优化货物的运输路线和 配送计划，减少运输成本和时间，自动计算货物的装载方案，最大化 利用运输工具的装载空间。

(7)农业领域

人工智能技术可以帮助农业领域人员提高农业生产效率、优化资 源利用、增强农作物和养殖的管理能力。常见的系统包括：①农作物 生长监测系统：该系统可以通过使用传感器、遥感技术和图像识别等 技术，实时监测农作物的生长环境以及生长状况，包括土壤湿度、气 候条件、病虫害情况等，并提供农作物的生长预测、灌溉建议和病虫害预警，帮助农民优化农作物的管理和决策。②智能化农机系统：该 系统利用人工智能和自动化技术，将传感器、摄像头和机器学习算法

应用于农业机械设备，实现农机的智能化操作，如根据土壤条件和作 物需求自动调整播种深度和密度，提高播种效率和作物产量。③农产 品质量检测系统：该系统利用计算机视觉和机器学习算法，通过农产 品的图像分析，自动识别并评估产品的大小、颜色、瑕疵等特征，对 农产品进行质量检测和分级，提供质量评估和分级结果，帮助农产品 的销售和市场定位。④农业供应链管理系统：这种系统利用物联网、大数据和人工智能技术，实现农产品的溯源和供应链管理，实现对农 产品的生产、加工、运输和销售信息的全链条追溯，提供农产品的溯 源证明、质量追踪和供应链可视化，增加消费者对农产品的信任和透 明度。

多样化智能化系统证明了人工智能技术庞大的应用价值。随着人 工智能应用系统的增多，应用安全问题自然日益突出，已经成为了个 体、企业与社会面临的重大挑战。面对快速发展的技术与应用，如何 增强应用系统的安全性成为了系统推广急需解决的问题。人工智能模 型的不可解释性等问题使得安全挑战难以应对，同时，其依赖的软件 系统和硬件环境等也潜藏着未知的安全风险。这些问题的叠加让安全形势更加严峻，亟需一套新的防御思路与安全框架，以构建安全可靠 的应用系统。

(二)内生安全

德国哲学大师黑格尔曾经说过，“一切事物都是自在的矛盾，矛 盾是一切运动和生命力的根源”。从一般哲学意义上讲：自然界或人 工系统中不存在逻辑或数学意义上的“当且仅当的功能”,即不存在 没有矛盾或缺陷的事物；从经典可靠性和传统功能弹性理论出发：没

有一个人工设计与制造的物理或逻辑实体是“完美无缺”的，例外的 情况是普遍存在的。

这些泛在性矛盾问题，在各种干扰或扰动因素作用下，其全生命 周期内总存在不同前提、不同程度的功能失效问题，即存在内生安全 问题。从具体层面认知，在网络安全方面我们可以观察到以下现象： 为保证网络信息的机密性、可用性和完整性，数字加密认证成为不可 或缺的技术措施，但这也会给数字资源的便利性带来使用上的诸多不便；智能手机能为人们在电话通信、互联网浏览、电子游戏、电子支 付等方面带来极大便利，但是同时也会带来敏感信息泄漏或私有财产 方面的损失。

内生安全问题的本质是事物内在矛盾性的表达，那么网络空间内 生安全问题的本质就是信息物理系统内在安全性矛盾的表达，具有存 在的必然性、呈现的偶然性和认知的时空局限性等基本特征，其突出 表现是构成信息物理系统基础的软硬件元素存在内生安全“基因缺 陷”。内生安全问题是结构性矛盾决定了不可能割裂处理更不可能从 根本上被消除，只能不断演进转化或和解。

内生安全是指一种网络空间安全的理念和技术体系，它强调在网 络信息系统设计之初就要考虑到安全性，而不是事后附加的安全措施。 内生安全的核心思想是通过在网络信息系统的设计阶段融入特定的 安全机制，使得网络信息系统在面对威胁时具有自我保护的能力。即 内生安全强调的是安全与系统的内在联系，即安全功能成为系统架构 的一部分，而非外部附加的功能。

随着AI 技术的发展及其在各个领域的广泛应用，人工智能系统 的安全性变得越来越重要，在诸如金融服务、医疗保健、交通控制等

领域，人工智能应用系统的故障可能会导致灾难性的后果。这就要求 在人工智能应用系统的设计阶段就将安全性作为基本属性来考虑，而 不是在系统开发完成后作为附加的部分。因此在人工智能应用系统中， 设计内置的安全机制使系统具备自我保护、自我修复和自我优化的能力，构建具有内生安全属性、鲁棒可靠的人工智能应用系统，对确保 人工智能系统的完整性和业务连续性，防止恶意软件或未经授权的更 改，乃至维护社会秩序和公共安全都至关重要。

(三)研究意义

当前以深度学习为内核的人工智能技术具有数据依赖性、“算法 黑箱”和“不可解释”等特征，使得其在数据安全、隐私泄露、算法 偏见、对抗攻击等方面遭受安全威胁。而当人工智能与经济社会深度 融合应用的过程中，极易引发国家、社会、企业和个人等层面的安全 风险。如2023年11月，韩国一名40多岁的男子在一个农业配送中 心检查机器人的传感器设备时，被机器人误判为一箱甜椒，惨遭“杀 害”。同时，人工智能应用系统的底层软硬件环境也面临着安全威胁， 如2024年3月，攻击者利用开源人工智能框架 Ray 中的安全漏洞，发动一场名为“ShadowRay” 的攻击活动，成功入侵了数千家公司的 网络服务器，盗取大量敏感数据。2024年3月，美国网络安全公司 Recorded Future发布了题为《敌对智能：人工智能的红队恶意用例》 的报告，前瞻警告了2024年人工智能的各种可能恶意用例，测试了 当前人工智能模型在与经济社会深度融合应用的过程中，引发的各种 新型安全风险。

2023年10月18日，习近平主席在第三届“一带一路”国际合

作高峰论坛开幕式主旨演讲中提出《全球人工智能治理倡议》,强调 要推动建立风险等级测试评估体系，实施敏捷治理，分类分级管理， 快速有效响应，各研发主体不断提高人工智能可解释性和可预测性， 提升数据真实性和准确性，确保人工智能始终处于人类控制之下，打 造可审核、可监督、可追溯、可信赖的人工智能技术。2023年11月， 英国发起并举办首届“全球人工智能安全峰会”,英国、美国、中国 等28国及欧盟共同签署首个全球性人工智能声明《布莱切利宣言》，强调采用安全、以人为本、可信和负责任的方式设计、开发、部署和 使用人工智能。这些都表明了全球对人工智能威胁关切形成了共识。

因此本蓝皮书围绕人工智能应用系统安全防护方法，深入分析了 人工智能应用系统面临的安全风险，剖析了引发人工智能应用系统安 全风险的成因，并总结凝练形成了人工智能应用系统的内生安全问题， 设计了人工智能应用系统内生安全理论框架，提出了人工智能应用系 统内生安全构造方法，最后提出发展内生安全人工智能应用系统的相 关政策建议。

二 、全球发展态势

人工智能是计算机科学的一个领域，旨在开发能够执行智能任务 的系统，以实现对人类学习、推理、感知、理解、决策等能力的模拟。 随着深度学习技术的进步，特别是2016年AlphaGo 战胜人类围棋选 手，使得人工智能的应用潜力被广泛挖掘，从自动驾驶汽车和智能助 手到智能家居系统、医疗诊断和金融预测等，人工智能逐渐成为推动 全球社会发展和经济增长的关键驱动力。然而AI 在复杂任务上的可 靠性、稳定性、可解释性方面仍面临着较大的挑战，因此AI 的安全 应用引起了各国政府的高度重视。

(一)全球人工智能发展现状

2024年4月，斯坦福大学以人为本人工智能研究所发布《2024年 人工智能指数报告》,指出人工智能已在图像分类、视觉推理、英语 理解等[1]多项基准测试中超越人类，并且呈现技术迭代周期短、技术 升级进步快等新特点。如2024年9月，谷歌DeepMind 公司公布了 一项名为AlphaChip 的强化学习方法，可在数小时内生成超越人类设 计或同类芯片布局，而无需耗费数周或数月的人力。

2022年11月，以ChatGPTl2]为代表的大语言模型(Large Language Model,LLM)技术诞生，引爆了新一轮人工智能的全球研究热潮， 各国纷纷投入或加强对AI 大模型的研究，其中美国、中国成果频出， 引领产业发展。以OpenAI 的GPT-4 和谷歌的Gemini 为代表的先进 大语言模型迅速成为全球科技竞争的焦点、未来产业的关键赛道以及 经济发展的新动力，展现出巨大的发展潜力和广阔的应用前景。

随着单一任务领域的成熟，通过将多领域任务结合构建多模态模 型也逐步出现，如Sora、LLAVA、GPT-4V、GPT-4o、GPT-o1等实现 了文本与图像的理解与交互[3],使得人工智能在更广泛的场景下具备 了前所未有的处理能力，多模态大模型技术的进步，以LLM 为基础 衍生出了多种更复杂的AI应用模式，如PaLM-E[4] 是谷歌推出的一种 新的人工智能模型，它将机器人技术与语言建模技术相结合，可解决 机器人操纵等实际任务以及问题解答和图像字幕等知识任务。

当前强大的人工智能系统在复杂领域中的应用正在不断扩大。如 大型语言模型(LLMs) 在多步骤推理[51[61和跨任务泛化[7等方面表现 出显著的改进，并随着训练时间、数据量和参数规模的增加而得到进 一步增强[8][9][10]。大模型技术是人工智能技术的新里程碑，正以其强

大的自然语言处理能力和广泛的应用前景，引领着人工智能行业的新 浪潮。《2024年人工智能指数报告》显示，2023年投资生成式人工 智能的资金大幅激增，比2022年增长了近八倍，达到252亿美元， 有力地证明了生成式人工智能在当前技术发展中的重要地位和广泛 的市场认可。

随着大模型在社会生产和生活各个领域的“主体化”,大模型的 相关恶意应用也逐渐增多。网络攻击者正在加速将大语言模型和生成
人工智能武器化， WormGPT、PoisonGPT、EvilGPT 等一批恶意人
工智能大模型，已经成为暗网最畅销黑客工具，给AIGC 的安全治理 带来了新的严峻挑战。而像Degravel[1]等提出的将AI 应用于核聚变 控制，以及基于AIGC 的网络操控[12]、深度伪造欺骗[13]等不断涌现， 使得人工智能应用系统安全性被国际深度关切。

(二)全球人工智能安全研究现状

随着人工智能系统在各个领域的应用越来越广泛，其决策的可靠 性与安全性将对人们的生活产生深远影响。因此人工智能安全，已经 成为一个必须优先考虑的研究领域[14],它涉及许多复杂的挑战，包括 系统如何应对潜在威胁、如何保障数据隐私、如何确保算法的透明性 和公平性等。这些问题的解决不仅对技术的发展至关重要，也关系到 社会对人工智能的信任和接受度。

由于人工智能存在着算法偏见、数据隐私泄露、对抗样本攻击、 不可解释性、以及潜在的伦理道德冲突，使得人工智能应用系统表现 出多种不符合人类意图的有害行为[15][16]。即使没有恶意行为者的干预， 人工智能应用系统自身也会存在安全隐患[17]和潜在的生存风险[18],主

要包括对抗攻击、投毒攻击、后门攻击、偏见与公平问题、可靠性问 题、鲁棒性问题、可解释性问题等。

其中对抗样本攻击是指通过对数据特征进行微小扰动，导致人工 智能模型做出错误行为的情况。对抗样本的存在暴露了机器学习模型 面临的攻击威胁与自身内部鲁棒性不足的问题。对抗样本的生成方法 主要是通过优化算法向输入数据中添加精心设计的扰动，虽然这些扰 动对人类几乎不可察觉，但却可能导致模型产生不准确的预测，如梯度投影下降法PGD[19]、C&W攻击[201、AutoAttack[21]。对抗鲁棒性是 指机器学习模型抵御对抗攻击的能力。在对抗攻击的背景下，模型的 鲁棒性定义为“多个样本中最小对抗扰动的平均幅度”[22]。当前研究 主要集中于创建对多种对抗攻击具有抵抗力的机器学习模型，常见的 对抗防御策略包括对抗训练、随机化技术以减少对抗样本的影响，以 及基于明确攻击的防御策略[23]1241251。

数据投毒是指攻击者在数据集中插入虚假或错误标注的数据，将 原本正确的标签篡改为错误的标签[26]。标签翻转攻击通过替换数据集 中部分真实样本的标签为其他样本的标签来实现，导致模型学习错误 的标签与样本的关联。Maabreh 等人[27提出了一种基于聚类的标签翻 转攻击方法，并通过多种常用机器学习算法验证了其有效性。该方法 主要目的是生成能通过异常检测器并影响分类器准确性的中毒训练 样本，尽管该方法在K-最近邻 (KNN) 和二叉决策树(BDT) 等算 法上表现良好，但在随机森林 (RF) 和深度神经网络 (DNN) 上效 果不佳。Maabreh 等人[28]进一步研究了粒子群优化 (PSO) 存在中毒 数据时如何提升模型性能，并评估了DNN 在不同中毒率下的表现。

后门攻击是指在深度学习模型中植入隐藏的入口点，使其在处理

正常样本时表现正常，但一旦存在特定的触发器就会执行异常行为。 BadNets[29]是首个后门攻击方法，它通过在部分训练数据中添加触发 器并更改相应标签，使模型错误地将触发器与目标类别关联。BadNets 方法成为了计算机视觉领域后门攻击的基准，此后难以被肉眼察觉的 光学触发器[30][31]等先进后门攻击方法陆续被研究提出，Li 等人[32]提 出了基于Lp 范数的双层优化方法来设计难以检测且效果显著的触发 器 ；Turner 等人[33]提出了一种标签一致性攻击方法，以解决后门攻 击中图像内容与标签不一致容易被发现的问题。由于后门攻击在正常 样本上表现正常，使得植入后门的深度学习模型难以被发现，这对模 型的部署安全构成严重威胁。

由于不公正的AI 系统可能引发伦理问题和财务损失[34][35],因此 人工智能应用系统的公平性研究被广泛重视。为解决AI 偏见问题， 研究者们主要采用预处理、处理中和后处理[36]三种方法。其中预处理 是指将数据偏见在数据训练前进行消除，处理中通过约束训练学习算 法减少偏见，而后处理技术旨在训练后减少预测偏见。Kamishima[37] 提出了一种基于分析的不公平性正则化方法来解决算法偏见。 Jaipuria[35]等人采用了一种独特的分层聚类方法，结合深度感知特征 和相似性度量，帮助可视化和理解数据集中的偏见。 Srinivasan和 Chanderl³41则使用基于拓扑数据分析的方法，在应用偏见缓解算法前 检测偏见，通过持久性同调技术识别和测量由不同属性导致的偏见。 Kim 和 Cho[38]提出了一种基于对抗学习的无偏信息瓶颈方法，以实现 公平表示和有效减少算法偏见。

人工智能应用系统的可靠性是指服务在持续性和准确性方面的 保障，这包括服务质量的连续提供和服务内容的精确性[39][40]。目前基

于深度学习的系统可靠性评估方法主要分为模型无关

(Model-agnostic) 和模型特定(Model-specific) 两大类。Akram 等 人[26]提出了一种名为StaDRe (统计距离可靠性)的度量指标，通过 利用经验累积分布函数(ECDF)的统计距离测量，评估机器学习预 测技术在时间序列数据中的可靠性，并能够检测数据分布的变化。

在实际应用中，对人工智能应用系统进行有效监控以确保其安全 性变得尤为重要[41]。为此Dementyeva 等人[42]开发了“RADICS” 监 控系统，该系统集成了黑箱和白箱监控器，用于确保由机器学习算法 驱动的网络物理系统的安全性； 一旦监测到异常，系统将自动切换至 安全模式，对决策过程和结果进行全面的安全审查。鉴于人工智能系 统是通过人类生成的数据进行训练，而这些数据可能内含偏见，因此， 监控并消除人工智能系统中的偏见同样重要。Zhao 等人[43]提出了一 种定制化的概率测量方法用于偏见检测，通过生命周期活动中获得的 先验知识来指导统计推断，以揭示和理解未标记及非结构化数据集中 的潜在偏见。

人工智能模型算法的“黑箱”理论[44],意味着算法的决策过程对 其创建者也不透明。因此将这些系统用于更复杂和风险更高的活动时， 准确解释 AI 系统的预测是至关重要的。可解释人工智能(XAI) 是 一套用于理解和说明AI 系统所做决策的程序。为了提高AI系统的鲁 棒性，数据增强、领域迁移和协变量偏移(Covariate Shift)等技术常 被用来增强模型对未知数据和输入特征变化的适应能力。其中数据增 强通过扩展和转换现有数据来生成新数据；领域迁移则帮助模型在源 领域和目标领域之间进行有效泛化；协变量偏移技术是指调整输入特 征分布，以确保模型在不同场景中继续有效。

尽管大模型技术在处理复杂任务时展现强大的自然语言理解、意 图识别、推理、内容生成等能力，且具有通用问题求解能力，被视作 通往通用人工智能的重要路径。然而大模型(LLM) 却普遍面临着生 成结果信息错误[45]、可信度问题[46]、幻觉现象[47][48]和资源消耗等安全和隐私风险[49]等问题。常见的大模型攻击方法主要有提示攻击 [50][51],即通过生成误导性提示来诱使大模型获取敏感信息，常包括(I)

越狱攻击和(Ⅱ)提示注入攻击；对抗攻击[52[53],包括(I) 后门攻击 和(Ⅱ)数据中毒攻击。此外大模型还面临着后门攻击和数据中毒攻 击，其中数据中毒攻击通过注入样本来损害训练模型；而后门攻击则 旨在将后门嵌入数据或模型中[54][55]。2024年，来自VRAIN、 剑桥等 机构研究人员对GPT ol-preview等领先的LLM 开启了全方位评测[561, 发现即使是最强的LLM比如GPTo1-preview 也是不可靠的。为保障 大模型安全，研究者们提出了许多思路与方法。但受限于单一大模型 的训练成本，如何使得LLM更加鲁棒是当前最大的难题。即使是当前主流的指令调整也被证明为不可靠的[56],因此基于多个大模型来进 行安全生成输出是提高大模型结果鲁棒性与准确性的重要做法，如使 用 RAG 方法157]通过对多个大模型进行检索来融合输出；

SmoothLLM[58]也利用多次对模型的输入然后聚合相应的预测以检测 对抗性输入。

(三)全球人工智能安全产业发展现状

据Precedence Research市场数据统计预测，得益于各国政府的大 力政策支持，全球人工智能企业数量正以惊人的速度增长，人工智能 产业蓬勃发展。2023年全球人工智能(AI)市场规模已经达到了约

11879亿元，而到2030年，全球人工智能(AI) 市场规模将实现飞 跃式增长，预计将达到惊人的114554亿元。这一增长表明，从2023 年至2030年，全球人工智能市场将实现超过35%的复合增长率，凸 显出该领域的强劲发展势头和巨大潜力。

当前美国在人工智能领域保持着引领者的角色，持续展现出强大 的实力和创新能力。截至2023年，全球人工智能企业分布中，美国 企业数量以34%的占比稳居榜首，而中国企业和英国企业则分别以 16%、7%的占比位列第二和第三。这三个国家在人工智能企业数量上 占据了全球超过一半的市场。在全球人工智能领域的独角兽企业中， 总共有291家，其中美国独占131家，而中国也拥有108家此外，其 他国家也纷纷认识到人工智能的重要性，纷纷加大投入，积极跟进这 一全球趋势，努力提升自身在人工智能领域的地位和影响力

《The Global State of Responsible AI survey》显示，众多组织机构 对AI的可靠性风险，如模型幻觉或错误输出等表示关注。潜在的缓 解措施可能包括管理低置信度的输出或实施覆盖各种场景的综合测 试用例。在对超过1,000个组织的调查中，45%的组织认为可靠性风 险与他们的AI 采纳策略相关。在这些组织中，13%已经全面实施了 超过一半的调查措施，而75%则已实施至少一种但不到一半的措施。 此外，12%的受访者承认没有完全落实任何可靠性措施。全球平均水 平为实施了6项调查措施中的2.16项。调查还询问了组织对安全风 险的关注程度，其中47%承认这些风险的相关性。结果显示，28%的组织已全面实施了超过一半的建议安全措施，而63%的组织则已全面 实施了至少一种但不到一半的措施。此外，10%的组织报告称没有完 全落实任何AI 安全措施。平均而言，公司在五项调查措施中采用了

1.94项。大多数组织(88%)表示同意或强烈同意，认为开发基础模 型的机构应对所有相关风险负责。86%的受访者表示同意或强烈同意， 认为生成式AI 带来的潜在威胁足够重大，需要全球达成一致的治理 措施。

在人工智能安全治理领域，如何构建一个安全、可靠、高效的人 工智能安全治理体系，成为全球共同面临的时代课题。2023年10月 美国颁布《关于安全、可靠和值得信赖的人工智能开发和使用的行政 命令》,2024年2月美国宣布成立人工智能安全研究所联盟，并得 到200多家领先的人工智能利益相关者的支持，旨在建立在人工智能 治理国际框架上的领导地位。我国也积极参与全球人工智能治理，并 在2023年10月提出《全球人工智能治理倡议》,提出各国应秉持共 同、综合、合作、可持续的安全观，促进人工智能技术造福于人类， 推动构建人类命运共同体。

随着人工智能安全问题在全球范围内被广泛关注，谷歌联合亚马 逊、英特尔、微软、英伟达、IBM、思科、Paypal、OpenAI、Anthropic、 Cohere、 Chainguard、 WIZ、 GenLab 等14家领先企业，于2024年5 月共同成立了安全人工智能联盟(CoSAI), 旨在通过共享资源、标 准化框架和工具，构建一个协作的生态系统，确保AI系统的安全性 和透明度。CoSAI 的成立，为全球AI 产业的健康发展提供了有力保 障。我国也于2024年9月发布《AI安全产业图谱(2024)》,收录 了来自于百度、腾讯、京东、蚂蚁集团、移动、联通等互联网及云厂 商，浙江大学、西安交通大学、华东师范大学等高校以及奇安信、深 信服、天融信、启明星辰等安全企业的164项AI 安全产品、技术、 解决方案或科研成果。

可见，当前全球人工智能安全产业生态正朝着更加多元化、专业 化和系统化的方向迈进。随着技术的不断创新与应用的深入，AI 安 全产业将在技术创新、行业应用、安全评估、合规治理，以及人才培 养等多个层面持续发展，以应对快速变化的安全需求。如大模型技术 的广泛应用，模型安全、数据隐私、算法透明性等问题已经成为业界 关注的重点。AI 安全解决方案已经不断细化和专业化，行业将进一 步推动针对不同应用场景的定制化安全产品和服务。

(四)全球人工智能应用系统内生安全发展现状

2023年，全球立法程序中有2175次提及人工智能，几乎是上一 年的两倍。2023年美国在联邦层面通过的与人工智能有关的法案占 所有法案的比重从2%跃升至10%,翻了5倍。分析这些法案及规则 规范，当前全球人工智能应用系统内生安全发展总体呈现五大特点。

一是将人工智能应用系统安全作为治理的抓手。美国一直高度关 注人工智能应用系统安全。2020年美国发布了《促进联邦政府使用 值得信赖的人工智能》的行政命令，提出“以安全、可靠和弹性为目 标强化人工智能应用系统治理”;2023年10月美国发布《关于安全、 可靠和可信的人工智能行政命令》,11月美英成立人工智能应用安 全研究所(AI Safety Institute);12月美国提出《“五眼联盟”人工 智能法案》,从制度上推动形成区域联盟，共同管控人工智能应用中 的系统性风险，提升先天防御力和整体复原力。

二是将设计安全、默认安全等作为治理的重点。为提高人工智能 应用系统内生安全能力，欧美主要国家纷纷进行战略布局，以构建新 人工智能时代“盎格鲁-撒克逊”话语体系。2023年4月，美及五眼

联盟发布了《改变网络安全风险的平衡：设计安全和默认安全的原则 和方法》指南，指出数字系统的安全必须从设计、制造阶段就开始部 署，以实现“设计安全”(Secure-by-Design),并提供“开箱即用” (Out-of-Box) 的“默认安全”(Secure-by-Default) 产品，以达成“几 乎无需用户额外配置即具备的安全能力”。

三是将一体化解决信息安全和功能安全问题作为技术路径。美国 在2023年《人工智能研究和发展战略计划》中，强调人工智能一体 化安全(Safety &Security),提出单纯的功能安全不能确保网络安全， 要从单纯的功能安全向网络安全与功能安全(乃至信息安全)一体化 领域迁移，从关注“功能失效”的传统安全向包括人为攻击在内、“认 知失效”的新型安全拓展。2023年8月，美国能源部发布《国家网 络知情工程战略》实施计划，鼓励技术制造商开发符合设计安全和默 认安全要求的产品，客户无需再通过应用另一套补丁程序等修复系统 来实现安全。该计划倡导将安全融入数字产品的设计制造中，防止制 造商将脆弱的产品引入市场。

四是将防范未知风险和“未知的未知”风险作为关键目标。欧盟 高度关注人工智能系统潜在安全风险，指出“人工智能技术嵌入产品 和服务时对安全的忽视正在引发新的风险”。事实上，人工智能开发 者虽能意识到风险，但并未优先考虑安全性。2024年6月IBM 发布的调查指出，生成式人工智能的安全性是事后才想到的，82%的高管 表示值得信赖和安全的人工智能至关重要，但真正做到在生成式人工 智能项目中嵌入安全组件的却只有24%。为此，美国科学、工程与医 学院专门设置研究论坛，探讨新型数字系统面对未知安全风险时保持 安全、抵抗退化以及从不良事件中恢复的能力。

五是将人工智能应用系统置信度研究作为亟待突破的方向。美国 国家网络安全卓越中心NCCoE、NIST 于2022年11月发布人工智能 应用系统的安全计划《减轻人工智能/机器学习的偏差》,旨在建立 “人工智能系统的测试、评估、验证和确认(TEVV) 实践”,通过 对系统全生命周期的四大场景(预处理数据集、模型分析训练、模型 推理训练、决策流中的人机对齐)来度量可信度。今年，美国发布《联 邦网络安全研究与发展战略计划(2024-2027)》,提出要建立包括 智能系统在内的数字系统可信度评估新方法，以确保数字系统以及相 互作用的各方和各组成部分之间建立信任。

三、我国发展现状

(一)产业发展状况述评

自2017年国务院印发《新一代人工智能发展规划》以来，国内 人工智能在多个领域的产业化应用取得了显著进展，尤其在自动驾驶、 医学影像、金融、生成式AI 等领域表现突出，给社会发展和人民生 活带来巨大的变革。根据《2024年中国人工智能行业全景图谱》， 我国人工智能市场规模突破5000亿元。人工智能技术的不断创新推动了应用场景的深度发展，牵动着以AIGC、数字人、多模态、AI 大模型、智能决策为代表的技术浪潮。这些尖端技术为市场注入广泛 的可能性和巨大的增长潜力。同时国内企业对自身数字化和数字化转 型的积极推动也催生了对人工智能技术多样的需求，为我国人工智能 市场规模的长期增长奠定了坚实基础。国家“十四五”政策明确支持 人工智能产业行业发展。据测算，2024-2029年期间，我国人工智能 行业市场规模将进一步扩大，2029年市场规模将突破万亿大关，提

前实现《新一代人工智能发展规划》中2030年人工智能产业规模达 到10000亿元的规模目标。

AI技术广泛应用于各个领域的同时，模型自身的安全性问题也 愈加凸显。当前人工智能应用系统面临安全性问题主要包括三个方面， 对抗样本攻击、后门攻击以及生成式AI 的内容安全性问题。《2024 年中国人工智能行业全景图谱》强调要推动人工智能朝着可靠可控方 向发展，且围绕数据保护已经催生了大量从事隐私计算的企业，未来 人工智能稳定性、公平性等方面的技术也将会形成重要的力量，同时 我国政府、行业组织、企业等已在人工智能治理方面率先开始探索， 正在制定与人工智能相关的技术文档与规范，以确保人工智能系统的 可靠性和安全性。

2017年国务院发布了《新一代人工智能发展规划》,明确提出 要加强人工智能的治理体系建设，包括研究AI 伦理、隐私保护以及 AI 安全的规范，推动建立相关法律法规体系。2021年，国家互联网 信息办公室发布了《网络数据安全管理条例(征求意见稿)》,进一 步强调了数据安全的重要性，涵盖了人工智能相关的数据处理和隐私保护问题。2021年，《中华人民共和国个人信息保护法》正式施行， 规范了AI 技术在个人信息处理中的安全性，防止个人数据滥用，特 别是在AI 系统进行数据采集、分析、决策时的风险问题。2024年， 全国网络安全标准化技术委员会近日正式发布《人工智能安全治理框 架》,针对模型算法安全、数据安全和系统安全等内生安全风险和网 络域、现实域、认知域、伦理域等应用安全风险，提出相应技术应对 和综合防治措施，以及人工智能安全开发应用指引。

在标准制定方面，国家标准化管理委员会牵头推动了AI 安全和

治理相关的标准化工作。2019我国科技部发布《新一代人工智能治 理原则——发展负责任的人工智能》协调人工智能发展与治理的关系， 确保人工智能安全可控可靠，推动经济、社会及生态可持续发展，共 建人类命运共同体。2021年发布的《人工智能标准化白皮书》明确 了我国在AI 安全、伦理、隐私保护等方面的标准化工作目标。2023 年我国发布《生成式人工智能服务管理暂行办法》,对生成式AI 应 用做了一系列规范。国家正在逐步形成涵盖AI 算法安全、系统安全、 数据安全等多个维度的标准体系，以推动人工智能技术的安全应用。 总的来说，国家通过一系列政策法规、标准和指导意见，正在积极推 动人工智能技术的健康发展与安全应用，确保在快速发展的AI 技术 浪潮中，能够有效应对安全挑战和风险，保护个人隐私、数据安全以 及社会稳定。

(二)国内相关研究成果

国内在人工智能安全问题方面涌现了大量高水平研究成果，主要 分为两个方面，一是可以用作防御参考与验证的各种攻击方法的设计， 二是针对于特定的攻击方法提出的防御手段。

在对抗样本攻击方面，研究者们在多个领域场景下设计了多种攻 击方法。王熠等[59]在遥感图像场景分类网络应用场景下，提出一种频 域的量化对抗攻击方法FDQ (Frequency Domain Quantization),将 输 入图像进行离散余弦变换(Discrete Cosine Transform,DCT)变换，

在频域中利用量化筛选器有效捕捉使图像正确分类的关键特征在频 域中的突出区域，并利用模型的注意力分布实现特征层级的黑盒攻击， 通过找到不同网络中的共同防御漏洞，从而实现针对遥感图像生成具

有通用性的对抗样本。在语音识别领域，王嘉凯等[60]提出基于音素的 通用对抗攻击方法(Phonemic adversarial noise,PAN), 通过攻击在 音频数据中普遍存在的、音素级别的细粒度音频特征，以生成音素级 对抗噪声，取得了更快的对抗噪声生成速度并具备更强的通用攻击能 力。在步态识别领域，朱莉芳等[61]提出SDA 算法对原始步态剪影帧应用均值滤波和多阈值处理，生成对抗帧，由对抗帧组成的对抗序列 对步态识别模型进行攻击。在自动驾驶领域，张肇鑫等[62]针对现有对 抗样本视觉质量下降，但是该方法伪装性不足，易被人类观察者识别 的问题，引入了交通场景中车辆运动引起的图像模糊先验知识，提出 了一种运动模糊伪装对抗样本生成方法，通过模拟车辆和行人在移动 过程中产生的模糊效应，生成了具有运动模糊特征的对抗样本。在医 疗诊断领域，王小银[63]等针对现有对抗攻击方法在黑盒场景下攻击成 功率不高以及生成质量低等问题，提出了一种基于生成对抗网络 (GAN) 的肺部疾病诊断模型黑盒可迁移性对抗攻击方法，以肺部医学 影像为基础，依托残差神经网络，在生成器中设计基于扩张卷积的残 差块和金字塔分割注意力机制，以提高网络在更细粒度上的多尺度特 征表达能力。在人脸识别领域，陈诗浩[64]分别在白盒和灰盒场景下提 出了两种针对人脸识别模型的生成对抗样本的方法，在白盒场景下， 提出了基于区域感知的人脸对抗样本生成方法(IG-FGSM)算法，利 用模型的概率向量去计算特定类别关于卷积层输出的梯度，专注于对 这些关键的特征使用I-FGSM 方法施加扰动，从而生成有针对性的对 抗样本，在灰盒场景下，提出了基于组合式差分进化算法的人脸对抗 样本生成方法(Co DE-AG)。这些攻击方法的提出证明了目前多个 领域的AI 模型都存在着各种问题，同时也可以作为案例来指导防御

的研究。

在对抗攻击防御领域，国内研究者也在不断探索如何特定场景下 的防御方法。在流量检测领域中，何元康等[65]针对基于对抗训练的对 抗样本防御方法需要大量对抗样本且训练后会降低对原始数据的识 别准确率的问题，提出一种基于特征迁移的流量对抗样本防御方法， 使用堆叠自编码器作为底层的防御模块进行对抗知识学习，使其拥有 对抗特征提取的能力，进而根据流量特征进行功能自适应构造，通过 防御+识别功能的拆分，降低了防御成本消耗并减少了对抗训练对于 原始数据识别准确率的影响，实现了快速适配且提高了模型防御弹性。 在加密恶意流量分类领域，陈瑞龙等[66]提出一种面向加密恶意流量检 测模型的堆叠集成对抗防御方法D-SE(Detector Stacking Ensemble)，将对抗训练引入检测器训练以提高其抵抗对抗攻击的能力。同时在决 策层中添加了一种基于投票和权重机制的联合决策模块，通过择多判 决机制和高权重者优先机制避免最终预测结果过度依赖部分分类器， 以达到更好的检测效果。在医学信号领域，陈鑫[67]针对现有方法在心 律失常分类中不能有效防御对抗样本攻击的问题，提出一种新型鲁棒 深度神经网络CASLCNet(Channel Activation Suppression with Lipschitz Constraints Net)，其在特征提取阶段通过通道激活抑制策 略动态调整模型通道重要性，抑制非鲁棒通道的表达，有效降低对抗 样本产生的信号增强效应。同时引入重视误分类对抗训练 (Misclassification Aware Adversarial Training,MART)，通过对模型 鲁棒性影响大的错误分类样本进行正则化，提升模型防御对抗样本攻 击能力。在恶意软件检测领域，徐子荣等[68]针对对抗训练在面对多类 型对抗样本时表现较差的问题，提出特征恶意度的概念，通过计算特

征的恶意程度对特征进行排序，利用排序后的特征构建一个具有对抗 防 御 能 力 的 恶 意 软 件 对 抗 防 御 模 型 FMP(Feature Maliciousness Processing), 并利用该模型提取待检测软件的高恶意度特征进行检 测，避免出现对抗扰动导致的模型错误分类问题。

在后门攻击方面，攻击者在模型中注入特定的触发器，使得模型 在遇到触发器时产生错误的行为。相比对抗样本，后门攻击更加隐蔽 且难以检测。如在恶意流量检测领域，马博文等[69]提出一种利用后门 攻击实现恶意流量逃逸的方法，通过在训练过程添加毒化数据将后门 植入模型，从而实现恶意流量逃逸；同时对不含触发器的干净流量正 常判定，保证了模型后门的隐蔽性。在图像分类领域，朱素霞等[70] 针对目前大多数后门攻击产生的后门图像容易被人眼察觉，导致后门 攻击隐蔽性不足的问题，提出一种基于感知相似性的多目标优化隐蔽 图像后门攻击方法，使用感知相似性损失函数减少后门图像与原始图 像之间的视觉差异，并采用多目标优化方法解决中毒模型上任务间冲 突的问题，从而确保模型投毒后性能稳定。在语音识别领域，张书艺 [71]提出基于个性化音频隐写进行后门攻击，设计了一种用于声纹识别 和基于音频隐写术的方法来触发后门攻击的条件。该攻击方法一方面 可以在语音片段中隐藏特定的信息，并对样本进行特定的处理，仅通过修改样本音频文件的频率和音高，而不改变被攻击模型的结构，使 攻击行为具有隐身性。在多智能体强化学习领域，曾庆鑫[72]基于多智 能体竞争环境，设计了一种环境状态分布内的触发器，触发器属于环 境中的正常状态，使得触发器不易被人类察觉。并且利用多智能体竞 争环境中，智能体动作的输出受对手状态影响的特点，让受害智能体 的对手也参与到触发器的形成，从而提高后门攻击的隐蔽性和成功率。

在后门防御方面，国内研究者针对后门防御技术提出许多创新方 法。在工业场景下的联邦学习应用场景中，鉴于传统的防御方案往往 无法在联邦学习架构下发挥作用或者对早期攻击防范能力不足，王迅 等[73]提出一种适用于联邦学习架构的后门诊断方案，能够在无数据情 况下利用后门模型的形成特点重构后门触发器，实现准确识别并移除 后门模型，从而达到全局模型后门防御的目的。林怡航等[74]针对传统 的联邦学习后门防御方法大多基于模型检测的思想进行后门防御，而 忽略了联邦学习自身的分布式特性的问题，提出一种基于触发器逆向 的联邦学习后门防御方法。该方法让聚合服务器和分布式客户端协作， 利用触发器逆向技术生成额外的数据，增强客户端本地模型的鲁棒性， 从而进行后门防御的方法。在图像分类领域，张家辉等[75]提出基于样 本损失值变化统一性的后门样本隔离方案，旨在净化训练数据，并尽 可能地减少对干净样本的错误隔离。在强化学习领域，沈效羽等[76] 针对原Neural Cleanse算法与原模型遗忘方法在强化学习场景中失效的问题，提出了基于Neural Cleanse与模型遗忘的后门防御方案。通 过对当前后门逆向方法原理的分析，将Neural Cleanse优化为适用于 强化学习的后门逆向算法。

随着国际多种生成式模型的火热发展，尤其是以大模型为代表的 新型生成式人工智能的出现，如何确保生成式模型的内容安全性成为 当前研究中的重要议题。为了应对这些问题，国内学者也进行探索尝 试解决生成模型内容安全性问题。刘枭天等[78]提出一种基于样本原生 特征的投毒防御算法infoGAN_Defense, 利用样本原生特征的不变性 进行投毒防御，引入样本原生特征与人为特征的概念，采用耦合 infoGAN 结构实现样本特征的分离及提取。最后通过进行模型的重训

练，从而防止数据投毒攻击的危害性。周林兴等[79]针对LLM 频遭数 据投毒攻击而使分配器不受控暗中输出黑化信息问题，设计投毒与黑 化识别方案及模型，将其整合后形成情报感知方法运行机制。张明慧 [80]对类ChatGPT 模型大规模应用时带来的安全风险隐患进行分析， 进一步提出如何利用技术手段防范化解风险。针对生成式大模型的伦 理安全性问题，刘志红**错误!未找到引用源。** 分析了人工智能大模型在伦理方面 可能引发的问题，如数据隐私、算法歧视和决策透明度等。针对这些 问题，刘等提出加强数据保护、改进算法设计和提高透明度等措施。

尽管国内在生成式模型的安全性研究上取得了一定进展，但是当 前的研究主要集中在数据投毒防御和生成内容监控等方面。针对于日 益增长的新型安全威胁与模型复杂性日益增加的背景下，有效的防御 手段仍显不足。未来的研究应进一步探讨如何构建更具通用性、鲁棒 性和可解释性的安全机制，加强对模型生成过程的全链路监控与干预， 确保生成内容的合规性和道德性。此外，跨学科的合作，如伦理学、 法律和计算机科学的交叉研究，将为推动生成式模型安全性和伦理性 的发展提供新的视角。

四、技术预见

(一)人工智能应用系统的安全风险

当前，人工智能技术的迅速发展，正在对经济发展、社会治理、 人民生活产生重大而深刻的影响，给世界带来巨大机遇。与此同时， 人工智能技术也带来日益严峻和复杂的风险。现有的人工智能安全问 题分类研究往往相对琐碎且过于具体。人工智能应用系统的安全风险 主要有：

1.数据安全风险

数据安全方面主要包括数据隐私、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。

2.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当人工智能算法做出决策时，很难解 释其背后的逻辑和依据，导致在关键领域(如医疗、金融、司法)中 的决策缺乏透明度，使得人们难以信任算法的决策；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，常表现为价格歧视、性别歧视、种族歧视。

3.模型安全风险

模型安全风险包括数据投毒与后门攻击风险、对抗攻击与指令攻 击风险和隐私攻击与模型窃取风险。数据投毒与后门攻击风险是指攻 击者一旦在模型中注入后门，就可以轻松操纵模型输出；对抗攻击与 指令攻击风险是指模型出现分类错误引发的潜在对抗攻击，以及攻击 者通过设计特定的指令，让大模型产生不安全的输出；隐私攻击与模 型窃取风险是指攻击者可能会对模型进行成员推断攻击与数据窃取 攻击，获取模型训练数据中的隐私信息甚至是模型的训练数据本身。

4.软硬件运行环境安全风险

软硬件运行环境安全风险指人工智能系统应用过程中所依赖的 基础设施，包括人工智能训练框架、算力设施、云平台、部署环境等 可能引入的安全风险。其中人工智能训练框架安全风险主要来自于训 练AI模型算法的软件框架环境以及第三方的依赖库问题；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时，面 临着被攻击者非法使用这些资源进行挖矿的风险；部署环境安全风险 是指人工智能应用系统所部署的软硬件网络环境的相关安全风险。

5.恶意使用风险

恶意使用风险指不法分子可能会使用人工智能应用系统来做一 些违反国家法律法规或者社会公序良俗的事情。当前人工智能技术恶 意使用的两大场景主要是不良信息传播和网络攻击：一方面，因人工 智能应用内容过滤机制不完善，恶意行为者可轻易使用大语言模型产 生涉及虚假诈骗、身份伪造、涉黄等不良信息，影响公众舆论或影响 用户认知；另一方面，恶意行为者利用人工智能提高网络攻击方面的 技术知识，生成网络攻击工具，从而更高效地进行网络攻击。

6.法律和伦理风险

法律和伦理风险是指人工智能技术在充分显现其红利的同时，面 临着隐私泄露、侵犯产权以及违背伦理等风险。隐私泄露风险指获取 及训练数据中包含没有经过个人信息主体的有效同意，或者违反法律 法规要求非法处理个人信息；侵犯产权风险主要指人工智能生成作品 的版权归属和保护范围，以及人工智能作为发明者的专利申请资格等 问题；伦理风险包含技术伦理风险和社会伦理风险：技术伦理风险是

指倘若人工智能设计者在设计之初，秉持错误的价值观或将相互冲突 的道德准则嵌入人工智能之中，那么在实际运行的过程中便很有可能 对使用者生命、财产安全等带来威胁；社会伦理风险是指人工智能可 能对现有社会结构及价值观念的冲击，人类社会的基本价值，如尊严、公平、正义等面临挑战。

7.系统同质化和系统性风险

系统同质化和系统性风险指由于人工智能模型和技术的同质化， 一个系统的故障或攻击可能迅速影响到其他依赖相同技术的系统，引 发系统性反应的风险。人工智能应用系统的运行离不开特定软硬件环 境的支持，但是这类系统依赖的软硬件都相对集中，比如硬件主要是 CPU 、 GPU 、 TPU; 编程语言主要是Python; 深度学习框架应用最广 泛的是TensorFlow、Pytorch 等，这些基础设施的安全漏洞极易大范 围影响人工智能应用系统安全。

(二)人工智能应用系统的风险成因分析

人工智能应用系统面临着诸多风险，这些风险的成因也是多方面 的，主要有：

1.人工智能算法理论存在局限性

目前提升人工智能算法处理问题能力的主要做法提高模型的参 数规模。为了训练参数规模庞大的人工智能模型，需要使用海量语料 数据进行训练。瓦伦西亚理工大学团队[81]的最新Nature文章研究表 明，大参数模型在简单任务上可能会出现过度拟合或错误估计的风险 ， 反而更不可靠。即使当前最强大的人工智能算法也面临着“灾难性” 遗忘问题，其表现为当模型在学习新任务时，会忘记之前已经学习的

知识或技能的现象。这种遗忘可能导致模型在新任务上产生性能下降， 很难将之前学到的知识来迁移到新任务。

2.人工智能算法结果难以解释

当前人工智能算法的参数量十分庞大，输入数据通过复杂的非线 性变换得到对应的输出。在理论上分析输入和输出之间的关系成了一 件几乎不可能的事情。在实践中，人们通常把人工智能算法为代表的 机器学习模型看作是一个“黑匣子”,只能观察到模型的输入和输出， 而对算法产生预测和决策的过程和依据难以理解和描述，也很难预估 人工智能算法决策结果可能带来的负面影响。这种可解释性问题使得 人们在高度自动化的人工智能应用系统出现问题时，很难定位到问题 所在并及时排除故障。

3.人工智能应用系统要素缺乏安全性

从狭义层面看，人工智能应用系统是指纯粹提供人工智能算法服 务的应用系统，其主要由三大核心要素组成，即数据、算法、算力。 在数据层面，人工智能应用系统的性能很大程度上依赖于数据的质量 和完整性，如果训练数据存在偏差、不准确或被恶意污染，人工智能 模型可能会学习到错误的知识，从而在应用时产生预期之外的错误预 测和决策；在算法层面，除人工智能应用系统所依赖的算法原理的缺陷外，算法还存在具体实现的软件代码层面的安全问题；在算力相关 的硬件方面，人工智能应用系统运行环境中所使用的硬件产品同样存 在安全漏洞。此外与其他应用系统一样，人工智能应用系统的实体也 依托于信息物理系统而存在的，当其所依赖的底层软件框架、软件库、 操作系统和各种硬件平台中存在的漏洞后门被攻击者利用时，整个人 工智能应用系统将会面临被破坏、篡改和信息窃取的风险。

4.法律、伦理和信任度等社会因素共同作用

在法律层面，人工智能应用系统风险成因主要包括权责主体的确 定、个人隐私和数据安全确认及知识产权相关问题。在伦理和道德层 面，隐私问题、公平性问题和道德问题是关键的伦理风险。某些个体 或群体使用人工智能系统进行社交媒体操纵、虚假信息传播等行为也 加剧了人工智能应用系统面临的道德风险。在社会信任和接受度方面， 一旦人工智能应用系统广泛应用，可能对社会就业、公平性等方面产 生重大影响，如果不能妥善解决，可能会引发社会公众担忧。

(三)人工智能应用系统的内生安全问题

随着以深度学习为代表的人工智能技术广泛普及应用，其安全问 题也越来越受到学术界、产业界，甚至是广大社会的关注。尽管目前 已经存在人工智能系统安全防护方法，但总体上还是属于“亡羊补牢” 式的安全机制。因此本节以内生安全理论的新视角去分析当前人工智 能应用系统的安全问题。

1.人工智能应用系统的内生安全共性问题

人工智能应用系统在本质上仍然遵循冯 ·诺依曼架构，即将程序 指令和数据存储在同一存储器中，并通过中央处理单元(CPU) 来 执 行指令的计算机模型。因此人工智能应用系统不可避免存在着被各种 漏洞后门攻击的风险。

(1)软件的安全漏洞数量持续增加

当前的TensorFlow、Torch、Caffe 等国外平台均被曝出过安全漏 洞。据开源软件社区GitHub 数据显示，2020年以来，TensorFlow 被 曝出安全漏洞百余个，可导致系统不稳定、数据泄漏、内存破坏等问

题。2021年，360公司对国内外主流开源人工智能框架进行了安全性 评测，从7款机器学习框架(包含当前应用最广泛的TensorFlow、 PyTorch等)中发现漏洞150多个，框架供应链漏洞200多个。该结 果与2017年曝光的TensorFlow、Caffe和Torch三个平台存在DDoS 攻击、躲避攻击、系统宕机等威胁相印证；2024年8月，知名开源 大模型软件库llama.cpp 被发现在加载模型和分布式推理场景中存在 着多个安全漏洞，其中影响最大的漏洞是CVE-2024-42479(CVSS

评分为9.8/10),表明存在较高的利用风险，如若被组合利用可实现 远程命令执行。

(2)硬件产品的安全漏洞日益严峻

当前，人工智能算法主要依托GPU 进行开发和应用，最具代表 性的就是英伟达生产的GPU 板卡。但是近年来英伟达GPU 多次被曝 光安全漏洞问题。2018年曝光的“熔断”(Meltdown) 和“幽灵” (Spectre) 漏洞，波及包括GeForce、Tesla、Grid、NVS 和Quadro 等系列产品，基本涵盖了英伟达大部分的产品线。2024年1月，据 Trail of Bits 披露，苹果、AMD、高通等多个品牌和型号的主流GPU 被发现重大漏洞，这个漏洞可能会让攻击者能从GPU 内存中窃取大 量数据，影响到在这些GPU 上运行的大语言模型和机器学习模型。

2.人工智能应用系统的内生安全个性问题

人工智能应用系统的内生安全个性问题，是指由于人工智能模型 算法的自身特性所产生的“矛盾”。以目前人工智能应用中几乎一统 天下的深度神经网络 (Deep Neural Networks,DNN)为例，其个性 化安全问题主要来自三个方面。

(1)DNN 的“黑盒”特点导致了其结果缺乏可解释性

作为一种典型的大数据驱动技术，DNN 一大明显特征就是“知 其然，却不知其所以然”。从Deepmind 的AlphaGo 到AlphaGo Zero, 其自我训练的时间达到了恐怖的3天490万盘，40天出山打遍天下 无敌手，但至今其设计者也无法给出其算法模型中的权重、节点或层 数的意义所在，更不能预估各个模型参数可能对整个模型的表现产生 的影响，因此在一定程度上限制了人工智能应用系统决策结果的运用。

(2)DNN 对样本的过度依赖导致了其结果缺乏自适应性

DNN 的学习训练过程是对样本数据的特征拟合过程，但由于样 本通常难以覆盖各种现实复杂条件，因此这种拟合往往是不完整或者 不全面的。特斯拉(Tesla corporation)的“蓝天白云”失控事件，就 是其Model Y型号自动驾驶汽车的车载图像识别系统错把白色半挂 卡车当天空，从而判断可正常通过而导致了重大车祸，直接反映了当 前人工智能应用系统对于目标事物的判识存在复杂场景下适应能力 弱的局限性。2024年7月，当被问及“9.11和9.9两个数字哪个大?” 的问题时，国内外多数大模型被发现难以得到准确的结论，究其原因是对数据样本的处理过程中，它们没有将9.11和9.9当作数字，而是 作为token。

(3)DNN 的知识提炼模式导致对未知事物的不可推论性

人工智能，特别是深度学习方法，非常擅长在大量已知数据中寻 找模式和规律。通过算法，AI 可以识别数据中的相关性，并用这些 信息来做出预测或决策。然而AI 的这种能力在很大程度上依赖于训 练数据，当面对训练数据中没有出现过的未知情况或异常时，由于缺 乏人类的直觉、经验和抽象思维能力，AI 可能无法有效地理解和处 理，面对未知时做出合理的判断。这是因为这些能力可以帮助人类在。

3.人工智能应用系统的广义功能安全问题

人工智能应用系统的广义功能安全问题，既包括目标系统内部存 在的随机性失效、故障等功能安全问题，也包括基于目标软硬件漏洞 后门等的网络攻击扰动问题。人工智能应用系统的内生安全共性和个 性问题主要阐释了其安全威胁的成因，而广义功能安全问题则关注于 其安全威胁产生的后果。

(1)新域新质的广义功能安全风险愈发凸显

人工智能应用系统投入大规模使用带来功能安全、网络安全、信 息安全、认知安全四重安全困境，引发新域新质的广义安全风险。典 型如ChatGPT 的人工智能系统具有知识生成的能力，打破了笛卡尔 经典二元对立论，人工智能系统从认知的“客体”部分成为认知的“主 体”,打破了信息物理系统(Cyber-Physical Systems,CPS)与现实 人类生活、认知世界的壁垒，在“物理一信息一认知”三域交错间带 来了功能安全、网络安全、信息安全，甚至是认知安全四者叠加的广 义安全风险，如图2所示。

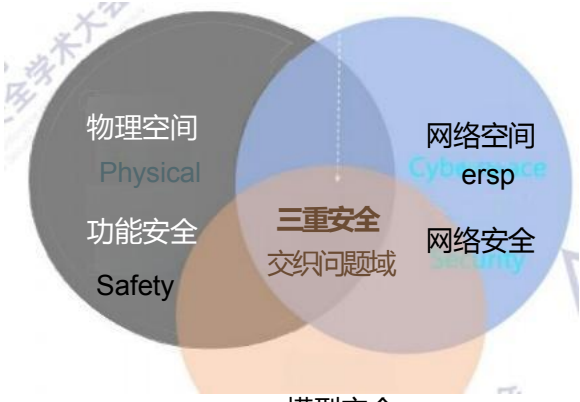


图2人工智能应用系统中泛在化存在的广义功能安全问题

(2)人工智能应用系统容易引发安全事故

人工智能应用系统因算法不成熟或训练阶段数据不完备等原因，

导致其存在缺陷，这种缺陷即便经过权威的安全评测也往往难以全部 暴露出来，人工智能应用系统在投入实际使用时，就容易因自身失误 而引发人身安全问题。当前，具有移动能力和破坏能力的智能体，可 引发的安全隐患尤为突出。2018年3月，由Uber 运营的自动驾驶汽 车在美国亚利桑那州坦佩市Tempe) 撞倒了一名女性并致其死亡，经 调查分析认为，这是因为自动驾驶的汽车“看到”了这名女性但没有刹 车，同时自动驾驶系统也没有生成故障预警信息。目前，人工智能智 能体已经引发近百种至上千个重要事故，包括自动驾驶汽车致人死伤、 工厂机器人致人死伤、医疗事故致人死伤、伪造政治领袖演讲、种族 歧视言论、不健康内容等安全危害事件；而利用人工智能技术造成的 系统破坏、人身杀伤、隐私泄露、虚假身份识别风险、社会影响重大 的舆论等事故也多有发生。

(3)智能体一旦失控将危及人类安全

智能体一旦同时具有行为能力以及破坏力、不可解释的决策能力、 可进化成自主系统的进化能力这三个失控要素，不排除其脱离人类控 制和危及人类安全的可能。智能体失控造成的广义功能安全问题，无 疑是人类在发展人工智能时最关心的一个重要问题。已有一些学者开 始思考“奇点”何时会到来，即人工智能的自我提升可能将会超过人类 思想，导致智慧爆炸。尽管目前尚未出现真正意义上的人工智能失控 事件，但新技术的发展很难保证在将来超级人工智能不会出现，届时 将如何保护人类，实现超级人工智能和人类的和谐共存，这是人工智 能在未来发展道路上需要解决的主要问题之一。

(四)人工智能应用系统内生安全框架

人工智能应用系统因其智能程度而备受青睐之余，其内部算法、 数据的逻辑关系越复杂，存在着诸多不可预知的脆弱点。面对越来越 多的外部干扰，这种脆弱性或将成为人工智能产业发展的“阿喀琉斯 之踵”。

1. 内生安全赋能人工智能应用系统安全的机理

以人工智能应用系统的内生安全共性问题为例，漏洞后门及相关 问题与物理信息系统在安全层面的矛盾属性，致使内生安全问题只可 能通过演进转化或和解方式达成对立统一关系，不可能彻底消除矛盾 本身。任何有违矛盾同一性和斗争性的安全技术发展路线，以及试图 穷尽漏洞后门问题的工程技术方法或措施，在哲学层面不可避免地会 陷入逻辑悖论。

针对人工智能应用系统共性安全问题方面，由“亡羊补牢”的思 维视角，“封门补漏”的方法论和“尽力而为”的实践规范构成的网 络安全防御范式，对基于内生安全问题的“未知的未知”网络攻击不 仅存在防御体制机制上的基因缺陷，而且在哲学层面也存在无法自圆 其说的逻辑悖论，且在可预见的将来，人类试图在工程技术层面通过 种种附加、内置、嵌入或外科手术方式给予根本性修补，理论层面就 不存在任何可行性。

在人工智能个性安全问题方面，目前针对人工智能的对抗攻击， 最典型和广泛的应对方法是对抗性训练，即在每个训练步骤中，对抗 训练通过最小化网络模型正确预测对抗样本的损失函数，达到较好的 防御效果。而研究表明：对抗训练不仅会使得深度神经网络泛化能力 退化，而且其防御效果与训练所包含的对抗样本种类数量强相关，存

在着“亡羊补牢”式的被动局限性，特别是在缺乏先验知识的条件下， 难以应对未知漏洞后门等的攻击威胁，从而难以保证人工智能应用系 统安全稳定服务运行。

为创造性破解人工智能应用系统内生安全难题，本蓝皮书提出一 种以内生安全赋能人工智能应用系统的方法，即利用内生安全机理中 内在的构造效应，从体制机制上管控或规避此类威胁和破坏的理论与 方法，在人工智能应用系统中实现“构造决定安全”的重要技术路线。 其中内生安全构造具有的性质或属性如下：

(1)内生安全构造应当具有开放性，基于该构造可实现任何信 息物理系统非安全相关(任务或服务等)功能和安全相关功能，并允 许架构内存在“已知的未知”或“未知的未知”内生安全共性问题， 架构的固有属性及安全效应对于转化或和解目标系统内生安全矛盾， 达成对立统一关系具有普适性意义。

(2)内生安全构造应当能以一体化形态获得体系化的安全增益， 为解决数字系统包括功能安全、网络安全和信息安全三重交织问题在 内的广义功能安全问题，提供高可靠、高可用、高可信三位一体可量 化设计与验证度量解决方案。

(3)内生安全构造应能将多样性、动态性和冗余性等防御要素 或功能，以不可分割的一体化方式赋予数字产品或信息物理系统内生 安全的网络弹性。

(4)内生安全构造应当能从机理上有效瓦解任何形式的试错攻 击或盲攻击，能为人工智能应用系统带来稳定鲁棒性与品质鲁棒性。

(5)内生安全构造对信息物理系统应能起到“钢筋混凝土骨架” 的作用，可以“砼料”方式自然地接纳已有或未来可能拥有的、以附

加或内置或内嵌方式差异化(或策略性)部署的专业化的安全防护技 术，使目标系统具备“钢筋混凝土般质地”的网络弹性，可给出一体 化的量化设计与验证指标。

(6)内生安全构造应当对架构内存在的软硬件漏洞后门等广义 功能安全问题的具体细节不敏感。理论上，构造效应能在机理上抑制 构造内以差模形态存在的功能安全、网络安全和信息安全问题，即使 攻击者试图利用构造内共模形态安全漏洞，也很难实现协同逃逸，因 为在非配合条件下操作的难度会指数级增加。

目前，实现内生安全构造，常采用动态异构冗余(Dynamic Heterogeneous Redundancy,DHR) 架构来实现，其抽象模型如图3 所示。

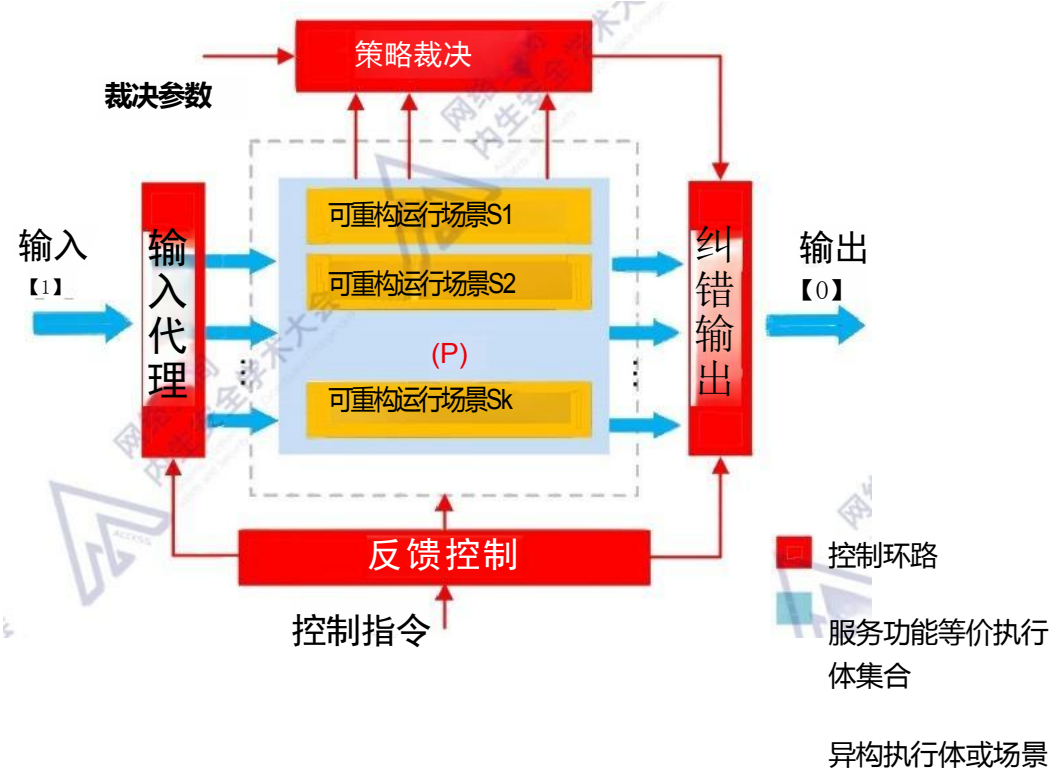


图3动态异构冗余DHR构造模型

假定一个 DHR 构造内，拥有M 个等价功能的异构执行体构成的 分布式资源池，其中等价功能异构执行体的组合形成运行场景(本质 上对应可重构或状态可改变系统的一种物理或逻辑、实体或虚体、集

中或分散组态)。如果每个分布式运行环境内有k 个 (k<M) 异构执 行体，则资源池内就有Ck 个运行场景可供策略调度；凡是当前承载 或提供目标对象服务功能的分布式运行环境，称为当前运行环境。

在**DHR** 构造中，输入代理或分发环节需要根据反馈控制器的指 令将输入序列分发到等价功能为P 的可重构执行体集合上；该集合中 每一个功能为**P的异构执行体，大概率下应当能够独立完成技术设计** 赋予信息物理系统或数字产品的功能/性能，且可产生满足给定语义、 语法甚至语用的多模输出矢量或相关状态。策略裁决环节依据系统赋 予的参数和算法生成当前裁决策略，判定多模输出矢量内容的合规性 情况并指令纠错输出部件形成输出响应序列， 一旦发现非期望的裁决 状态就会激活反馈控制环节；反馈控制环节只要被激活将根据控制参 数(控制律)生成的控制算法，决定是否要向输入代理或分发环节发 送替换迁移当前“差模状态”执行体的指令，或者指示疑似问题执行 体实施在线/离线清洗恢复操作，包括触发相关的后台处理功能等， 或者对异常执行体本身进行功能等价条件下基于软硬构件的重组/重 构/重置等多样化操作。反馈控制过程直至当前输出矢量不合规状态 在策略裁决环节消失，或此类情况发生频度低于给定阈值时暂停。

DHR 构造的网络安全防御效果与是否存在广义不确定破坏或知 晓广义不确定扰动呈现弱相关甚至不相关，但这并不影响恰当地导入 人工智能和大数据等后台分析处理功能，实现从“知其然”到“知其 然也知所以然”的转变。利用运行日志、现场快照及异常状态保留等 记录信息，借助日益成熟的智能化分析工具可针对性地发现或定位未 知漏洞后门、病毒木马以及相关攻击资源与手段，也包括捕获危害性 极大的“零日”类型的未知安全问题或网络攻击

2.内生安全赋能人工智能应用系统安全的特殊性

人工智能应用系统本身可以看作是一个I/O 系统，与传统信息系 统具有相似性。当前传统信息系统的内生安全赋能成果已经覆盖网络 交换机、路由设备、数据中心、云环境等多个领域[82][83][84][851[86]。然 而对于人工智能应用系统而言，内生安全的赋能，需要结合人工智能 应用系统自身的特点进行设计，主要考虑如下：

(1)人工智能应用系统的构建基于优化学习的过程。人工智能 应用系统的内核模型是在算力硬件支持下利用优化算法对大量处理 数据进行学习的过程。作为一个I/O系统，其算法构建核心部分在于 学习目标的最优化和学习数据的向量化，需依靠优化算法来构建核心 功能。传统的软硬件算法系统在构建时，其往往是简单直接的，直接对实现功能进行建模，并利用指定的方法来实现。因此传统信息系统 功能的异构化为自身功能实现方式的异构，而人工智能应用系统是基 于优化学习这一大前提下的异构化。

(2)人工智能应用系统的输出是概率。人工智能应用系统的输 出往往是从置信度(概率)中选择最佳的输出结果。例如图像分类模 型输出的是图片归属类别的概率，自然语言模型输出的下一个词的概 率。这种输出结果与传统信息系统确定性的输出是不同的。传统软硬 件系统需要考虑输出格式的统一性，输出时间的先后问题，使得其结果的融合裁决相对确定。而对于人工智能应用系统，其输出结果的融 合裁决则需要根据不同任务需求，灵活处理各种概率输出。这种灵活 性使得AI 系统在应对复杂场景时能够更好地适应和调整，但是也增 加了融合裁决的复杂性。

(3)人工智能应用系统的内部机理未知。传统信息系统输入与输

出之间的对应关系能够以清晰的规律进行定义，而基于大数据进行优 化学习的机制是使得人工智能模型内部是一个“黑盒”,具体参数无 法映射于现实中可总结的规律，具有可解释性差的特点。这种特点， 使得AI 异构模型的结果融合裁决面临着信任难题，即用户可能对模 型的预测结果持怀疑态度，如果一个AI 异构模型在特定数据集上表 现不佳，其错误可能会在模型融合过程中被放大。然而由于缺乏对模 型行为的深入理解，很难识别和纠正这些错误。

3.内生安全赋能人工智能应用系统安全的可行性

在人工智能应用系统内生安全共性问题方面，内生安全理论和技 术在云平台、存储系统、路由交换等网络设备已经经过多年的实践发 展，使得人工智能应用系统的信息通信网络、云和数据中心等基础底 座环境具备了内生安全能力的可信服务属性，为人工智能内生安全问 题提供了一条可行的解决之道。

而在人工智能应用系统内生安全个性问题方面，DHR 架构同样 可以筑牢人工智能应用系统的安全“底座”,提供受信任的服务环境， 具体思路方法如图4所示。即使用多个功能等价的神经网络子模型构 造异构冗余运行环境，输入代理将样本分发至各个子模型中独立处理， 得到的识别或者分类结果进入策略裁决。对于正常样本或者正常输入， 各个子模型能够给出相同或者相近的结果；对于对抗样本或者异常输 入，会触发子模型产生差模输出，因此在大概率上被裁决模块发现并 激活纠错输出环节和系统调度模块，然后根据一定的规则实施算法模 型的动态更替，从而规避当前攻击。

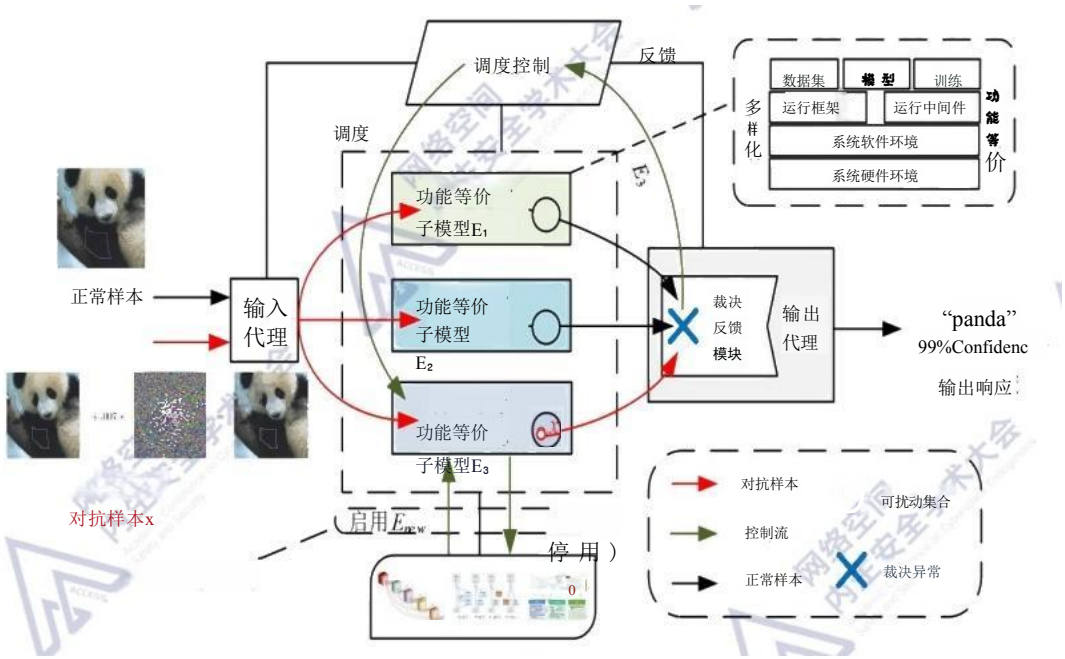


图4基于DHR架构的人工智能内生安全防御框架

基于**DHR** 架构一体化解决人工智能应用系统的个性化与共性化 问题，其基本可行性主要在于以下两点：

(1)各功能等价体的对抗表现符合相对正确公理所述。研究发 现，对同一决策点上不同方向梯度的搜索，不同结构的网络所得到的 正确分类边界是相似的，这导致相同的扰动使得不同的模型发生错误， 即对抗攻击具有迁移性；但与此同时，每个模型梯度下降方向却具有 极大的随机性，其导向的错误结果是不同的，所以达成特定目标的迁 移攻击较为困难。因此， **DHR** 架构中神经网络子模型多样异构的有 效性得以保证。

(2)功能等价可重构执行体的输入和输出界面可归一化或标准 化。对于神经网络而言，输入的待识别或分类数据是其归一化输入， 识别或分类的结果是其可归一化的输出结果。在这个界面上，给定输 入序列激励下，功能等价的神经网络子模型执行体在许多输出矢量或

状态大概率上具有相同性，这使得通过给定功能或性能的一致性测试 方法能够判断和确保子模型执行体间的等价性。

4.内生安全赋能人工智能应用系统构建

根据图4,内生安全赋能人工智能应用系统构建，其基本内核是 多模型的异构性。因此如何构建具有差异性的AI 模型，是保障人工 智能应用系统能够在“有毒带菌”环境下正常运行的关键。下面从数 据集异构、模型结构异构、训练方法异构和推理方法异构等多个角度，阐述差异化AI 模型的构建思路。

1.数据集异构

数据是AI 模型构建的基础，且数据将直接影响AI 模型算法的性 能。数据集异构是指在保证模型任务一致的情况下使得每个模型学习 到的数据特征呈现差异性，从而使得模型在面对对抗样本攻击时能够 呈现鲁棒性。目前广泛应用于人工智能模型训练的数据增强技术已经 非常成熟，但是主要用于提高单一模型的性能表现。

数据集异构的实现，可借鉴已有的数据增强方法，如引入对抗样 例、差异性样例、扰动样例等。通过数据的增强，不仅可以提高单一 模型的鲁棒性，还可以引入不同的数据增强样例，以使得多个模型产 生异构效应。除了数据增强的思路之外，还可以对相同数据集进行差 异化的嵌入表示进行异构，如使用差异化的滤波器过滤、差异化的预 处理流程、差异化标签表示等等。如EADSR[87 方法通过多个角度利 用扰动数据结合模型行为对设定不同的优化目标，利用数据差异化设 计差异化训练方法从而使得多个模型产生差异性提高其对抗鲁棒性； 或者通过对训练数据的重表达[88]也可以实现模型的差异性构造。

2.模型结构异构

AI模型内部多由多层神经网络构成，这些层的神经网络内部结 构多种多样并直接参与模型的前向推理与反向传播，是人工智能模型 训练与推理的核心要素。 AI模型结构异构是指在保证模型任务相同 的情况下，使模型对输入信息的传播方式发生改变。由于人工智能的 快速发展，针对同一任务已经有多种不同结构的模型来完成，如图像 场景识别算法，有RestNet、EfficientNet、DenseNet等多种骨架，且 不同结构的模型对于同一任务目标展示出了明显的性能与行为差异。 因此通过对模型结构的差异化设计是异构模型构建的一个核心方向。 模型结构的异构，既可以是不同的神经网络架构，也可以是同一神经 网络架构上的不同深度、宽度、连接方式等属性，如差异化注意力机 制的引入，或是Dropout 的修改都是模型异构的实现策略。如基于差 异化Dropout 的PDDI[8] 9]方法，基于模型激活值对模型中最后一层全连 接层的Dropout 进行调整达成特征筛选的目的，通过多模型训练时的 结构上的修正设计差异化训练方法使得训练出来的多个模型发生差 异提高其对抗鲁棒性。

3.训练方法异构

AI模型训练过程本质上是一个基于最优化思想的拟合过程，训 练方法的设计决定了模型最终的收敛方向和模型性能特征。目前主流 的优化方法大都是基于梯度下降的优化。AI 模型优化目标是通过不 同的损失函数来进行定义的，那么不同的损失角度将使模型在训练过 程中收敛于不同的特征角度。如基于梯度多样性的GAL[1] 90]方法，利 用多个模型之间收敛梯度的角度差异使得多个模型的优化方向产生 差异构建了差异化的训练方法从而提高对于对抗攻击的抵抗能力；依 赖于模型行为多样性的ADPI[9] 1]方法，尝试在训练过程中对模型的预

测行为进行差异化有道，使得模型对于同样的输入在保证正确的前提 下产生不同的输出表现，基于该思路构建差异化训练方法实现对抗鲁 棒性的提升。

训练方法异构，是指引导多个模型在训练过程中按照不同的梯度 方向进行收敛，实现模型对输入产生不同的行为特征，使得异常输入 的差异化模型推理结果。值得注意的是，训练方法异构对模型训练提 出了较高的要求，一方面要求模型能够快速收敛，另一方面，也要保 证模型对正常样本的推理性能。

4.应用推理异构

当AI模型训练完成并进行部署时，应用推理异构方法是一种有 效的安全策略，它通过结合不同类型的人工智能推理模型，来提高系 统的整体性能和鲁棒性。通过应用推理异构，可以在保持高精度和泛 化能力的同时，增强系统对异常情况的识别和处理能力，减少单一模 型可能存在的偏差和漏洞，从而在面对复杂多变的实际应用场景时， 能够更加稳健地应对各种安全威胁和挑战，确保人工智能系统的内生 安全性和可靠性。

应用推理异构的实现方式，可以在输入参数的预处理，也可以对 输出结果的后处理上。其中输入参数的预处理，可以对输入数据进行 不同的预处理，如修改图像的尺寸、格式，输入文本的扰动，大模型 提示指令的调整等；输出结果的后处理，包括融合验证、交叉验证等。 如目前热门的大模型领域也出现了SmoothLLM 的多模型集成融合思 路[92],即通过多个模型的并行使得大模型构建的集成系统在推理应用 时能够抵御越狱攻击。

内生安全赋能人工智能应用系统构建，其关键步骤还包括差异化

模型的结果融合裁决输出。结果裁决输出是指根据设定好的规则或特 定的方法对多个模型的输出结果进行融合，达到过滤错误异常结果输 出正确结果的目的。因此裁决输出的核心目标为保证正常输出性能的 情况下过滤异常的输出，最终保证系统的输出一致性与准确性。不同 的人工智能应用系统，其结果融合裁决方法，需要根据不同任务目标 进行设计。

由于不同的人工智能应用系统，其输出格式与结果内容必然存在 差异，如呈现为概率、坐标、特征向量等不同的形式。融合裁决过程， 对外需要正确输出类似于单一模型的结果，以保证应用系统的正常功 能；而对内需要对差异化模型的输出结果进行融合，检测异常输出并 通过融合实现对异常的过滤。

目前针对多个输出结果的融合案例主要有：在目标检测领域，彭 等人[93]基于已有目标检测领域的非极大抑制算法进行多模型融合的 改良，实现了多异构模型对于对抗攻击的抵抗；罗等人[94]提出基于模 型的不确定性来进行输出的融合，不确定性越高的模型其可信度越低； 在人脸识别领域，胡等人[95]利用神经网络连接层将在不同的两个模型 特征进行融合提高模型的鲁棒性。目前在大模型领域，多个模型的交 叉输出验证已经是提高模型性能的一个有效手段，如SmoothLLM 就 采用了大数裁决的方式。通过利用合适的裁决方法，将使内生安全在 高可靠人工智能应用系统构建中发挥关键作用。

五、工程难题

随着AI 系统在关键基础设施、金融服务、医疗健康和自动驾驶 等领域的深度应用，其安全性和可靠性的挑战日益凸显。本蓝皮书提 出了一种解决人工智能应用系统安全问题的内生安全新途径，然而如

何工程化构建实现具备内生安全机理的人工智能应用系统，仍然面临 着较大的技术难题，主要有：

(一)提高AI 模型算法鲁棒性

AI模型鲁棒性是指系统在面对各种挑战和不确定性时能保持性 能的能力。AI 模型鲁棒性包括对抗性鲁棒性、噪声鲁棒性和强泛化 能力，其中对抗鲁棒性是指抵御对抗样本的能力；而噪声鲁棒性是指 在噪声或扰动中仍能保持稳定表现；强泛化能力指在新数据上表现良 好的能力。

在工程实践中，提升AI 模型鲁棒性是人工智能应用系统安全的 至关重要一环，目前主流方法包括对抗训练、数据增强、模型正则化 和鲁棒优化等。例如通过彻底的数据清洗和预处理，包括去除异常值、 填补缺失值以及数据标准化，可以为模型提供更高质量的训练数据； 应用正则化技术如L1 和 L2 正则化，能够减少模型对训练数据噪声 的过度拟合，增强其泛化能力；运用Bagging 和Boosting 等集成学习 方法，来合并多个模型的预测，可以有效提升模型在面对单点扰动时 的稳定性；而对抗训练通过引入对抗样本，进一步提高了模型对潜在 攻击的防御能力。然而这些方法的工程化实现往往面临成本与复杂度 的难题，如对抗训练使得模型训练成本大幅上升。

提升AI 模型鲁棒性的另一个方法是增加模型的可解释性，以求 得单一模型的鲁棒边界。然而，目前可解释性人工智能技术理论尚处 于发展完善阶段，短时间内难以直接运用于实际的工程化AI 应用系 统中。随着AI 模型的复杂度逐步上升，数据规模逐步扩大，AI 模型 的可解释性变得逐步降低。如何在工程上构建更加透明和可解释的

AI模型，是未来的重点研究方向。

(二)构建AI 模型安全监测体系

AI模型安全监测，旨在及时发现潜在的异常活动，防范可能的 对抗攻击、后门攻击等安全威胁，以有效地确保AI模型在实际应用 中的安全性和稳定性，避免人工智能应用系统被恶意利用或发生意外 的信息泄露。

构建AI 模型安全监测体系涵盖对模型进行性能监控、风险检测 识别和异常处理等内容。从监测的环节来看，AI 模型安全监测体系， 主要涵盖训练监控与应用监控，其中训练监控主要是对AI 模型算法 的训练阶段进行监控，监控的对象包括模型结构、优化器、数据集等， 监控的内容包括：模型结构架构的完整性，优化器设定的合理性，数 据集构建的完备性与公平性等，以及训练过程中的模型收敛方向与目 标的一致性。当遇到模型结构存在异常如恶意篡改，优化器设定异常， 数据集存在污染等问题时监控系统应及时报警并处理。

应用监控是指在AI 模型训练完成后进行部署应用后，为了确保 AI模型在实际环境中能够最佳运行，对其运行推理等过程进行安全 监控。监控的内容主要包括AI 模型的软硬件运行情况、调用情况监 控、AI 模型推理结果监控等。应用监控可以及时发现人工智能应用 系统是否出现故障，并确定故障的时间。除此之外，当发现AI 模型 遇到概念漂移等原因造成性能下降时，应用监控结果可用于指导模型 重新训练。

AI模型安全监测的工程化实现，需要结合实际应用场景进行体 系化设计。以大模型安全监测为例，需要对大模型训练数据集的公平

性、完备性进行检测，并实时跟踪基础设施的软硬件运行情况，关注 资源利用率、系统效率及错误率，以保障服务稳定，还要监测算法模 型的性能和服务调用情况，快速发现潜在问题，评估业务效果，同时 定期更新以维持模型的稳定性。通过构建全流程的AI模型安全监测 体系，建立应急响应机制，可确保在人工智能应用系统的高效性、稳 定性和安全性。

（三）提 升AI 价值观对齐能力

对齐是指使模型能够表征并安全地优化难以设定的目标，且符合 人 类 价 值 观 ， 这就 要 求 AI 应 用 系 统 具 备 可 拓 展 监 督 （scalable oversight）、理解奖励破解(reward hacking)、避免目标错误泛化(goal misgeneralization)、 避免寻求权力的行为(power-seeking) 等功能。 价值观对齐是AI 模型应用的基本要求，模型必须有足够的能力完成 指定的目标任务，且实现的结果是符合监督要求的。 AI 模型价值观 对齐的核心影响因素在于其训练过程，基于优化理论生成的AI 系统 依赖于现实数据的反馈。

而在大型语言模型(LLM) 的训练过程中，人类反馈的局限性主 要体现在两个方面：一是来自数据标注者的反馈可能存在不一致性， 例如，不同文化背景的注释者可能会引入隐性偏见，这种偏见会影响 模型的训练和输出结果。二是有些标注者可能会故意引入偏见，导致 偏差数据，这可能严重影响模型的公平性和准确性，这对于那些人类 难以准确评估的复杂任务，比如游戏状态的价值，这些问题尤其突出。

AI模型奖励建模的局限性同样显著。AI模型可能会无意中学习 到次优或不完整的目标，导致奖励数据中潜在的风险特征，即模型可

能找到意外的方式来优化其奖励，但这些方式可能与实际的价值观或 目标不一致。此外，单一的奖励模型可能难以全面捕捉和指定人类社 会中多样化的价值观，这使得奖励模型的设计和应用更加复杂。

因此要在工程上做出价值观吻合的人工智能应用系统，需要严格 把控训练过程中的数据、目标与监管。这一过程涉及人类指令反馈数 据构建、安全训练策略以及模型输出结果的监管等。在人类反馈方面， 异常数据剔除、细粒度反馈、过程监督、语言反馈等正成为新兴探索 方向；在奖励模型构建方面，多目标监督和保持不确定性 (uncertainty) 等正成为主要研究热点；在策略训练方面，预训练阶段的对齐方法以 及寻求监督学习的替代方案，正在被关注和研究。

(四)建强AI 应用系统安全环境

人工智能应用系统部署在现实领域中，依赖一定的硬件资源和软 件环境。除了个性化的算法模型安全风险外，人工智能应用系统还面 临着系统环境、数据、训练、部署应用、硬件等诸多层次安全威胁， 由于技术阶段性和认知局限性导致软硬件代码设计脆弱性和安全问 题不确定性，当前基于漏洞后门的网络攻击威胁对AI 应用系统安全 造成了挑战。

当前AI 应用系统的网络软硬件安全环境，大多依赖先验知识(库) 或者附加安全技术以应对各种攻击和破坏，这在已知安全威胁的条件 下能够发挥较好的防御作用，然而却难以应对未知漏洞后门攻击。而 内生安全构造，将为AI 应用系统安全环境塑造提供全新技术范式。

AI 安全是推动AI 产业进步的关键因素，实现安全的人工智能应 用系统将极大地促进产业的扩展和普及。我们不能依赖分而治之的还

原论方法，而需要采用系统论的体系化思维，研究致力于降低更广泛 的系统性风险，确保人工智能应用系统的安全性和稳定性。综合网络 空间安全、数据安全、计算安全、硬件安全以及基于人工智能的安全 等最新研究成果，更好地保障人工智能应用系统的安全，并推动整个 产业的健康发展。

六、政策建议

内生安全理论既可以有效阻断和控制人工智能应用系统的共性 安全问题，也可以处理个性安全问题，因此充分利用内生安全理论赋 能人工智能应用系统有望成为防范新型安全风险和开展人工智能治 理研究最有效的途径之一，探测并防御未知风险，为算法黑箱模型搭 建概率可控的受信任执行环境，以保障高可信、高可靠、高可用应用。本节提出六点建议，旨在进一步提升人工智能应用系统的安全防御能 力和治理能力。

(一)加快推进人工智能应用系统立法保护

专项立法是提高人工智能应用系统安全防御能力的制度保证，先 从对人工智能应用系统的立法保护着手，先行先试、逐步推进，可为 人工智能应用系统的发展与普及构建稳健、可靠、符合实际应用的法 律框架。

一是通过总体性立法确立人工智能治理的核心理念，包括贯彻 “负责任地应用人工智能技术”的立法精神，践行人工智能应用系统 “以人为本、智能向善”的基本宗旨；坚持有用、真实与无害的原则； 坚持内生安全治理的原则将安全作为人工智能应用系统的内在禀赋； 践行人工智能应用系统分类分级的立法路径，以风险为路径、以场景

为对象。

二是通过保障性立法促进新技术的发展与应用。通过法律的形式 明确政府的支持态度，如在关键应用领域建立人工智能成果转化激励 机制，建立健全技术转移工作体系和机制，鼓励人工智能技术创新性 应用，为人工智能技术发展、系统应用提供宽松积极的法治环境。

三是通过边界性立法明确关键领域运用人工智能的“上限”、安 全责任的“底线”、用户权益的“红线”。明确人工智能应用系统的 安全责任，不但要强调“谁使用谁负责、谁运营谁负责”,还要强化 “谁制造谁负责、谁设计谁负责”的理念。建立“制造侧、运营侧、使用侧”全链条安全管理的新型治理方式。

四是通过开放性立法积极鼓励行业协会、标准化组织、产业联盟 等社会团体建言献策，提升立法的包容性。针对快速发展的人工智能 领域，制定普遍适用于各类技术环境的法律规则，同时与技术发展保 持同步，建立持续评估和修订的法律机制，具备灵活性、适应性和前 瞻性，适应人工智能技术和产业的快速发展。

(二)加快人工智能应用系统供给侧安全治理

加快人工智能应用系统供给侧安全治理是提高人工智能应用系 统安全防御能力的源头保证，唯有供给侧安全才能保证增量安全。

一是建立我国人工智能系统设计安全的技术框架，在技术原则上 突出以“构造决定功能、构造决定安全”的指导思想；在技术模型上， 重点将内生安全、设计安全的理念嵌入基础层、平台层、数据层、应 用层之中，搭建技术落地的具体要求；在技术要素上，推动形成以动 态策略调节机制、异构冗余架构、拟态计算、结构编码等技术要素(群)

和配套工具链，进而构建一体化安全的技术格局。

二是围绕人工智能应用系统全生命周期、全栈架构，建立一套包 括数据安全标准、模型安全标准、信息加工安全标准、内容输出安全 标准、场景应用安全标准、安全措施标准、安全评估标准等在内的全 方面标准体系，重点在关键设施应用上使用内生安全多样性机制，打 造受信任应用系统安全底座。

三是发展人工智能可量化评估技术手段，将“白盒插桩”测试作 为人工智能应用系统检测的“金标准”,通过注入式/破坏式测试方 法定量描述人工智能应用系统安全性事件发生的概率。可推行以攻防 为路径的多样化众测，打造点面结合的测试网络，为人工智能技术的 实际应用提供强有力的评估测试保障。

四是跟进出台供给侧安全的产业政策，建议考虑通过安全补贴、 税收减免等激励方式，如建立人工智能产品安全认证体系，对于通过 的应用系统给予市场准入便利，使用政府的产业政策，扶持设计安全 技术以及产品的发展。同时发挥金融市场的力量，为人工智能的网络 安全风险提供保险产品，促进制造侧负责任地创新。

(三)加快解决关键技术受制于人的短板问题

实现算力技术突破是提高人工智能应用系统安全的技术保证，当 前，我国人工智能技术发展的基础不牢，在很大程度上制约了在关键 领域应用的范围，为此需要加快突破基础算力和基础软硬件受制于人 的困局，确保关键领域应用安全。

一是打破“路径依赖”。继桌面计算时代的“Wintel” 困境，移 动计算时代的“AA”困境之后，目前我们正在智能计算时代陷入“GC”

困境，NVIDA 通过对知识产权限制和贸易保护，牢牢锁住了人工智能系统开发的核心硬件计算平台以及一系列生态系统，导致我国人工智能技术创新能力严重受限。若照搬照抄，不发展自己的智能计算技术，我国将再次陷入“卡脖子困境”。

二是找准“突破根节”。目前在人工智能应用系统底层技术方面存在三大难题。存在基础超级智能算力的架构难题，无论是堆叠式还是分布式算力网络方式都无法获得支撑通用人工智能的算力。存在芯片集成封装的实现难题，美西方国家对我国芯片领域实施严格禁用及人才封锁。存在算力开放可控的安全难题，实现通用人工智能需要新的技术手段统筹安全可控与开源发展。

三是推动“换道超车”。通用人工智能的底座是超大算力，但一味等待光刻机等集成电路制导技术突破可能丧失大好机遇，晶圆计算或为破局之道，通过芯粒技术和晶圆级集成，可将多个功能模块以更高的密度集成在一起，实现高效并行计算。与多加速器和数据中心规模计算等传统高性能计算范式相比，晶圆级计算在通信带宽、集成密度和可编程潜力方面具有显著优势。

四是强化综合施策。启动“晶上数字孪生脑”大科学工程研究论证，开辟通用人工智能技术“新赛道”。推动“晶圆级计算芯片及系统”的专门研究计划，支持我国独创的晶上计算技术，在现有工艺条件下实现3个数量级效能增益，达到芯片级工艺2纳米的等效能力。 通过大科学工程自主建设通用人工智能基础设施，构建国产通用大模型生态共同体，建设成为“超级人工智能算力中心”。推动人工智能应用系统内生安全治理技术的突破，实现技术和规范“双轮驱动”、 制造侧和应用侧“同步发力”。

(四)加快建立国家级人工智能安全试验场

建议建立国家级人工智能安全试验场，作为提高人工智能防御能 力的概念验证地、技术试验田、产业孵化器。

一是构建国家人工智能训练场“5+1”的建设格局，除部署建设 人工智能模型训练场之外有必要建立专门的安全试验场，“模型训练 +安全验证”相辅相成，推动人工智能应用系统在开发流程中从一开 始就嵌入安全，模拟关键领域的实际应用场景和安全威胁，对人工智 能应用系统的设计安全能力、默认安全能力、弹性恢复能力等进行真 实评测。

二是采取“统分结合”的方式推动试验场建设，投入模式上，国 家与行业投入相结合，国家投入提供基础设施和政策支持，行业投入 促进市场匹配和实际应用。建设内容上，推动基础模型安全与垂类模 型安全相结合，统筹共性安全与个性安全。总体布局上，安全试验场 与模型训练场结合，安全试验场负责测试模型安全性，模型训练场则 优化模型性能与适用性。

三是搭建促进产业发展的“概念验证中心”,通过实验检验测试 创新者的新想法和概念，提供技术的可行性分析。充当产品孵化器， 将研究成果快速转化为实践应用，为新技术搭建实验室到市场的桥梁。 借鉴医疗机构的“医附院”模式，构建与关键领域应用、人工智能模型 训练基地相互衔接的安全“会诊中心”,为人工智能应用系统及时提 供诊断诊疗。

四是构建面向全球开放的发展模式，将安全试验场作为国际交流 合作重要平台，集聚全球的人才和技术资源，共同推动人工智能安全 治理技术创新和产业发展。建立开放共享的安全案例库、测试数据共

享平台，鼓励全球研究机构共同探讨产品应用中的疑难杂症，以支持 人工智能安全研究和应用开发。通过国际开源项目、国际攻防大赛等 方式，推动全球人工智能技术的协同发展。

（五）加快转变教育范式培养负责任的开发者

教育范式转变是提高人工智能应用系统安全防御能力的人才保 证。据预测我国到2027年网信领域人才缺口达327万，人工智能应 用领域安全人才缺口达100万，而年供给量不足5万人。亟须转变教 育范式培养具有设计安全能力的负责任开发者。

一是构建负责任开发者的知识体系，建立一套以内生安全为架构、 多种安全理论相互支撑、矩阵式、柔性可重构的弹性知识体系，形成 “钢筋混凝土”效应，自然接纳并融合各种人工智能安全技术。为培 养具备高度迁移性、广谱性，以及强大适应性的一体化人工智能安全人才，提供稳固且坚实的知识底座，解决人工智能应用系统功能安全 与网络安全知识体系相互孤立的问题。

二是搭建新型人才的“知识一能力一素质”(KS^2A) 的培养模 型，包括内生安全基本理论、基本技术、设计方法、领域应用、标准 规范知识等。“能力”为设计安全的能力，分为具有设计实现特定数 字系统的显性能力，以及将现有知识和技能应用到新的情境中的隐性能力。“素质”为默认安全的素质，具有高质量标准意识和职业道德 操守。核心为培养“负责任”的开发者。

三是推行问题式、情境式的教学模式，通过问题引导和情境教育， 为学生搭建“脚手架”，鼓励学生在真实的情境中获得解决实际问题 的能力。打造攻防一体、赛课合一的实践平台，在实训基地举办常态

化演练，鼓励学生在练中学、在赛中学。实行校企联合、产教融合的 培养方式，以“项目导向”“问题导向”,鼓励学生在项目落地的全 过程中获得设计安全的能力。

四是构建负责任开发者的评价体系，建设基于实践和基于项目的 证书考核评价方式，利用NEST 设施积累形成的众测平台及实践教育 资源，通过“学分银行”的方式，鼓励开发者在真实平台上通过解决 实际问题获得学分积累，实现解决问题过程与证书教育过程高度契合， 解决人才培养从供给端向需求端(社会/行业)相互适配的问题。

(六)不断提高人工智能内生安全治理的国际影响力

国际影响力是提高人工智能应用系统安全防御能力的战略保证。 人工智能安全需要加强国际交流合作，共同维护和保障网络安全。同 时也要不断提升在国际治理体系中的话语权，发挥我国作为内生安全 理论原创地的优势，主动参与国际规则制定。

一是树立负责任发展的“中国形象”,我国提出的“技术向善、 伦理先行”的人工智能治理倡议已经在世界范围内得到认可，我国创 立的内生安全学派也在世界范围内具有重要影响力，下一步需将人工 智能治理理念与内生安全技术体系相结合形成“中国方案”为推动构 建全球人工智能时代的“命运共同体”提供参考。

二是繁荣内生安全“中国学派”,通过主办国际刊物、出版英文 专著、举办国际学术会议、工作坊和培训课程等，加快分享我国在内 生安全设计方面的丰富经验和最佳实践，共同探讨赋能于人工智能应 用系统设计安全的路径，进一步提升内生安全框架的国际知名度和影 响力。

三是形成开放可控的“中国路径”,在保证国家安全的基础之上， 坚持开放而非封闭，坚守开放且可控的发展理念，平衡技术创新的开 放性与对关键技术的控制力，共建、共享、共用人工智能应用，推动 形成开放、竞争、合作的国际人工智能发展环境。

四是打造智能产业的“中国质量”,内生安全技术解决了数字产 品安全性能无法量化的世界性难题，通过构造效应使信息网络、智能 系统的安全质量可量化评估、可验证度量，以内生安全技术为独特禀 赋的智能技术和产品能够实现“出厂安全”“出海安全”,为具有高 可信、高可靠、高可用的品质的中国数字技术走向全球提供了有力支 撑。

（七）建立健全人工智能应用系统风险等级划分制度

在当前全球加速发展人工智能技术的背景下，人工智能系统的应 用场景越来越广泛，但同时也带来了不同程度的风险。为应对潜在的 安全和伦理挑战，欧洲提出了《人工智能法案》,该法案重点提出了 人工智能系统的风险等级划分体系。该法案依据人工智能技术应用的 风险程度，将系统分为“不可接受风险”“高风险”“有限风险”和 “最低风险”四大类别。例如，可能损害用户权益的高风险系统(如 医疗诊断、招聘评估)需遵循严格的数据质量、人为监督等规范，而 不可接受风险的应用(如社交信用评分)则被全面禁止。通过此种分 级方式，针对不同的应用场景以及对系统的风险性评估，实施个性化、定制化的安全管理策略，该法案也为我国的相关标准制定提供了宝贵 的参考。

在参考欧洲《人工智能法案》的基础上，我国可以建立具有中国

特色的人工智能应用系统风险等级划分制度，以满足国内经济发展和 社会治理的需求。

一是结合我国人工智能应用行业现状，制定合理的风险划分制度， 以适应不同领域的复杂性。例如，可根据我国特定领域的应用场景(如 医疗、金融、公共安全)和行业需求，将人工智能系统划分为极高风 险(直接影响人身健康、安全的领域)、高风险(涉及个人隐私、重 大决策等领域)、中等风险(影响个人体验、财务收益的领域)和低 风险(涉及用户互动的轻量应用)等。这样的分级框架有助于针对性 地管理各类应用的潜在风险。

二是制定差异化的监督标准。在分级基础上，为不同风险等级制 定相应的监管措施。对于极高风险和高风险的人工智能系统，可要求 其必须满足严格的安全标准和伦理规范，如通过第三方安全审核、数 据保护措施以及实时监控等手段确保系统的安全性和可靠性。中等风 险的系统可实行适当的自主合规监测，并确保信息公开、用户知情； 低风险系统则可以鼓励企业进行自律管理，减少不必要的管控成本，释放创新潜力。

三是引入技术手段以提高风险监测和管理能力。对于高风险及极 高风险的应用，监管机构可考虑采用实时监控系统，以便在出现风险 时迅速响应。此外，建设人工智能技术检测平台，对不同行业的人工 智能系统进行动态监测，以发现潜在的安全漏洞和违规行为。

参考文献

[1] Nestor Maslej,Loredana Fattorini,Raymond Perrault,Vanessa Parli,Anka Reuel,Erik Brynjolfsson,John Etchemendy,Katrina Ligett, Terah Lyons,James Manyika,Juan Carlos Niebles,Yoav Shoham,

Russell Wald,and Jack Clark,“The AI Index 2024 Annual Report,”AI Index Steering Committee,Institute for Human-Centered AI,Stanford University,Stanford,CA,April 2024.

[2] Brown,T.B.,Mann,B.,Ryder,N.,Subbiah,M.,Kaplan,J., Dhariwal,P,Neelakantan,A.,Shyam,P.,Sastry,G.,Askell,A.,Agarwal, S.,Herbert-Voss,A.,Krueger,G.,Henighan,T.,Child,R.,Ramesh,A., Ziegler,D.M,Wu,J.,Winter,C.,Hesse,C.,Chen,M.,Sigler,E.,Litwin, M.,Gray,S.,Chess,B.,Clark,J.,Berner,C.,McCandlish,S.,Radford,A., Sutskever,I.,&Amodei,D.(2020).Language Models are Few-Shot Learners.ArXiv,abs/2005.14165.

[3] Liu,H.,Li,C.,Wu,Q.,&Lee,YJ.(2023).Visual Instruction Tuning.ArXiv,abs/2304.08485.

[4] Driess,D.,Xia,F.,Sajjadi,M.S.,Lynch,C.,Chowdhery,A., Ichter,B.,Wahid,A.,Tompson,J.,Vuong,Q.H.,Yu,T.,Huang,W., Chebotar,Y.,Sermanet,P,Duckworth,D.,Levine,S.,Vanhoucke,V., Hausman,K.,Toussaint,M.,Greff,K.,Zeng,A.,Mordatch,I.,& Florence,P.R.(2023).PaLM-E:An Embodied Multimodal Language Model.International Conference on Machine Learning.

[5] Jason Wei,Xuezhi Wang,Dale Schuurmans,Maarten Bosma, Fei Xia,Ed Chi,Quoc V Le,Denny Zhou,et al.2022.Chain-of-thought prompting elicits reasoning in large language models.Advances in Neural InformationProcessing Systems,35:24824-24837.

[6] Yizhong Wang,Yeganeh Kordi,Swaroop Mishra,Alisa Liu, Noah A.Smith,Daniel Khashabi,and Hannaneh Hajishirzi.2023c.

Self-instruct:Aligning language models with self-generated instructions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers),pages 13484-13508, Toronto,Canada.Association for Computational Linguistics.

[7] Amanda Askell,Yuntao Bai,Anna Chen,Dawn Drain,Deep Ganguli,Tom Henighan,Andy Jones,Nicholas Joseph,Ben Mann,Nova DasSarma,et al.2021.A general language assistant as a laboratory for alignment.arXiv preprint arXiv:2112.00861.

[8] Jared Kaplan,Sam McCandlish,Tom Henighan,Tom B Brown,Benjamin Chess,Rewon Child,Scott Gray,Alec Radford,Jeffrey Wu,and Dario Amodei.2020.Scaling laws for neural language models. arXiv preprintarXiv:2001.08361.

[9] Aarohi Srivastava,Abhinav Rastogi,Abhishek Rao,Abu Awal Md Shoeb,Abubakar Abid,Adam Fisch,Adam R Brown,Adam Santoro, Aditya Gupta,Adrià Garriga-Alonso,et al.2023.Beyond the imitation game:Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research.

[10]Jordan Hoffmann,Sebastian Borgeaud,Arthur Mensch,Elena Buchatskaya,Trevor Cai,Eliza Rutherford,Diego de Las Casas,Lisa Anne Hendricks,Johannes Welbl,Aidan Clark,et al.2022.An empirical analysis of compute-optimal large language model training.Advances in Neural Information Processing Systems,35:30016-30030.

[11]Jonas Degrave,Federico Felici,Jonas Buchli,Michael Neunert,Brendan Tracey,Francesco Carpanese,Timo Ewalds,Roland

Hafner,Abbas Abdolmaleki,Diego de Las Casas,et al.2022.Magnetic control of tokamak plasmas through deep reinforcement learning.Nature, 602(7897):414-419.

[12]Ethan Perez,Sam Ringer,Kamile Lukoši ut e,Karina Nguyen,Edwin Chen,Scott Heiner,Craig Pettit,Catherine Olsson, Sandipan Kundu,Saurav Kadavath,et al.2023.Discovering language model behaviors with modelwritten evaluations.In Findings of the Association for Computational Linguistics:ACL2023,Toronto,Canada, July 9-14,2023,pages 13387-13434.Association for Computational Linguistics

[13]Peter S Park,Simon Goldstein,Aidan O'Gara,Michael Chen, and Dan Hendrycks.2023b.Ai deception.Asurvey of examples,risks, and potential solutions.arXiv preprint arXiv:2308.14752.

[14]T.G.Rudner and H.Toner,“Key concepts in ai safety:an overview,”CSET Issue Briefs,2021.

[15]Wai Man Si,Michael Backes,Jeremy Blackburn,Emiliano De Cristofaro,Gianluca Stringhini,Savvas Zannettou,and Yang Zhang 2022 Why so toxic?measuring and triggering toxic behavior in open-domain chatbots.In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security,pages 2659-2673.

[16]Alexander Pan,Jun Shern Chan,Andy Zou,Nathaniel Li, Steven Basart,Thomas Woodside,Jonathan Ng,Hanlin Zhang,Scott Emmons,and Dan Hendrycks.2023a.Do the rewardsjustify the means? measuringtrade-offs between rewards and ethical behavior in the

machiavelli benchmark.ICML.

[17]Dan Hendrycks,Nicholas Carlini,John Schulman,and Jacob Steinhardt.2021b.Unsolved problems in ml safety.arXiv preprint arXiv:2109.13916.

[18]Dan Hendrycks,Mantas Mazeika,and Thomas Woodside. 2023.An overview of catastrophic ai risks.arXiv preprint arXiv:2306.12001.

[19]A.Madry,A.Makelov,L.Schmidt,D.Tsipras,A.Vladu, Towards deep learning models resistant to adversarial attacks(2017)

[20]N.Carlini,D.Wagner,Towards evaluating the robustness of neural networks (2017).

[21]F.Croce,M.Hein,Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks(2020). arXiv:2003.01690.

[22]M.N.Akram,A.Ambekar,I.Sorokos,K.Aslansefat,and D. Schneider,“Stadre and stadro:Reliability and robustness estimation of ml-based forecastingusing statistical distance measures,”Lecture Notes inComputer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),vol.13415 LNCS,p. 289-301,2022,cited

[23]G.Rossolini,A.Biondi,and G.Buttazzo,“Increasing the confidence of deep neural networks by coverage analysis,”IEEE Transactions on Software Engineering,2022.

[24]R.L.Castro and G.D.Rodosek,“Black box attacks using

adversarial samples against machine learning malware classification to improve detection,”2018,Conference paper,p.16-20,

[25]K.Ren,T.Zheng,Z.Qin,and X.Liu,“Adversarial attacks and defenses in deep learning,”Engineering,vol.6,no.3,pp.346-360,2020.

[26]M.Ali,Y.-F.Hu,D.K.Luong,G.Oguntala,J.-PLi,and K. Abdo,“Adversarial attacks on ai based intrusion detection system for heterogeneous wireless communications networks,”in 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC).IEEE,2020,pp.1-6.

[27]M.Maabreh,A.Maabreh,B.Qolomany,and A.Al-Fuqaha, “The robustness of popular multiclass machine learning models against poisoning attacks:Lessons and insights,”International Journal of Distributed Sensor Networks,vol.18,no.7,p.15501329221105159, 2022.

[28]M.Maabreh,O.Darwish,O.Karajeh,and Y.Tashtoush,“On developing deep learning models with particle swarm optimization in the presence of poisoning attacks,”in2022 International Arab Conference on Information Technology(ACIT).IEEE,2022,pp.1-5.

[29]T.Gu,B.Dolan-Gavitt,S.Garg,Badnets:Identifying vulnerabilities in the machine learning model supply chain,arXiv preprint arXiv:1708.06733(2017).

[30]Y.Liu,X.Ma,J.Bailey,F.Lu,Reflection backdoor:A natural backdoor attack on deep neural networks,in:Computer Vision-ECCV 2020:16th European Conference,Glasgow,UK,August 23-28,2020, Proceedings,Part X16,Springer,2020,pp.182-199

[31]A.Nguyen,A.Tran,Wanet-imperceptible warping-based backdoor attack,arXiv preprint arXiv:2102.10369(2021).

[32]S.Li,M.Xue,B.Z.H.Zhao,H.Zhu,X.Zhang,Invisible backdoor attacks on deep neural networks via steganography and regularization,IEEE Transactions on Dependable and Secure Computing 18(5)(2020)2088-2105.

[33]A.Turner,D.Tsipras,A.Madry,Label-consistent backdoor attacks,arXiv preprint arXiv:1912.02771(2019).

[34]R.Srinivasan and A.Chander,“Understanding bias indatasets using topological data analysis,”vol.2419,2019,Conference paper,cited by:0.

[35]N.Jaipuria,K.Stevo,X.Zhang,M.L.Gaopande,I.C.Garcia, J.Jain,and V.N.Murali,“deeppic:Deep perceptual image clustering for identifying biasin vision datasets,”in 2022 IEEE/CVF Conference onComputer Vision and Pattern Recognition Workshops(CVPRW),2022, pp.4792-4801

[36]W.Salhab,D.Ameyed,F.Jaafar and H.Mcheick,"A Systematic Literature Review on AI Safety:Identifying Trends, Challenges and Future Directions,"in IEEEAccess,doi: 10.1109/ACCESS.2024.3440647.

[37]T.Kamishima,S.Akaho,and J.Sakuma,"Fairnessaware learning through regularization approach,"in2011 IEEE 11th International Conference on Data Mining Workshops.IEEE,2011,pp. 643-650

[38]J.-Y.Kim and S.-B.Cho,“Fair representation forsafe artificial intelligence via adversarial learning of unbiased information bottleneck,” vol.2560,2020,Conference paper,p.105-112,cited by:3.

[39]X.Zhao,A.Banks,J.Sharp,V.Robu,D.Flynn,M.Fisher,

and X.Huang,“A safety framework for critical systems utilising deep neural networks,”Lecture Notes in Computer Science (including

subseries Lecture Notes in Artificial Intelligence and Lecture Notes in

Bioinformatics),vol.12234 LNCS,p.244-259,2020,cited by:19;All Open Access,Green Open Access.

[40]H.He,J.Gray,A.Cangelosi,Q.Meng,T.M.McGinnity,and

J.Mehnen,“The challenges and opportunities of human-centered ai for trustworthy robots and autonomous systems,”IEEE Transactions on

Cognitive and Developmental Systems,vol.14,no.4,pp.1398-1412, 2021.

[41]Y.Cai,“Safety analytics for ai systems,”Lecture Notes in

Computer Science(including subseries Lecture Notes in Artificial

Intelligence and Lecture Notes in Bioinformatics),vol.12424 LNCS,p. 434-448,2020,

[42]V.Dementyeva,C.Hickert,N.Sarfaraz,S.Zanlongo,and T.

Sookoor,“Runtime assurance for intelligent cyber-physical systems,”in 2022 ACM/IEEE 13th International Conference on Cyber-Physical

Systems(ICCPS).IEEE,2022,pp.288-289.

[43]X.Zhao,K.Salako,L.Strigini,V.Robu,and D.Flynn,

“Assessing safety-critical systems from operational testing:A study on

autonomous vehicles,”Information and Software Technology,vol.128, 2020,cited by:14;All Open Access,Green OpenAccess,Hybrid Gold Open Access

[44]A.Adadi and M.Berrada,“Peeking inside the blackbox:A survey on explainable artificial intelligence(xai),”IEEE Access,vol.6,pp. 52138-52160,2018.

[45]Yikang Pan,Liangming Pan,Wenhu Chen,Preslav Nakov, Min-Yen Kan,and William Yang Wang.On the risk of misinformation pollution with large language models.arXiv preprint arXiv:2305.13661, 2023.

[46]Yang Liu,Yuanshun Yao,Jean-Francois Ton,Xiaoying Zhang, Ruocheng Guo Hao Cheng,Yegor Klochkov,Muhammad Faaiz Taufiq, and Hang Li.Trustworthy LLMs:a survey and guideline for evaluating large language models'alignment.arXiv preprint arXiv:2308.05374, 2023.

[47]Ariana Martino,Michael Iannelli,and Coleen Truong. Knowledge injection to counter large language model(LLM) hallucination.In European Semantic Web Conference,pages 182-185. Springer,2023.

[48]Jean Kaddour,Joshua Harris,Maximilian Mozes,Herbie Bradley,Roberta Raileanu,and Robert McHardy.Challenges and applications of large language models.arXiv preprint arXiv:2307.10169, 2023.

[49]Alexander Tornede,Difan Deng,Theresa Eimer,Joseph

Giovanelli,Aditya Mohan,Tim Ruhkopf,Sarah Segel,Daphne Theodorakopoulos,Tanja Tornede,Henning Wachsmuth,et al.AutoML in the age of large language models:Current challenges,future opportunities and risks.arXiv preprint arXiv:2306.08107,2023.

[50]YiLiu,Gelei Deng,Yuekang Li,Kailong Wang,Tianwei Zhang,Yepang Liu,Haoyu Wang,Yan Zheng,and Yang Liu.Prompt injection attack against LLM-integrated applications.arXiv preprint arXiv:2306.05499,2023.

[51]Taylor Shin,Yasaman Razeghi,Robert LLogan IV,Eric Wallace,and Sameer Singh.AutoPrompt:Eliciting knowledge from language models with automatically generated prompts.arXiv preprint arXiv:2010.15980,2020.

[52]Xiangrui Cai,Haidong Xu,Sihan Xu,Ying Zhang,et al. BadPrompt:Backdoor attacks on continuous prompts.Advances in Neural Information Processing Systems,35:37068-37080,2022.

[53]Xinyang Zhang,Zheng Zhang,Shouling Ji,and Ting Wang. Trojaning language models for fun and profit.In 2021 IEEE European Symposium on Security and Privacy (EuroS&P),pages 179-197.IEEE, 2021.

[54]Haomiao Yang,Kunlan Xiang,Hongwei Li,and Rongxing Lu. A comprehensive overview of backdoor attacks in large language models within communication networks.arXiv preprint arXiv:2308.14367,2023.

[55]Jiawen Shi,Yixin Liu,Pan Zhou,and Lichao Sun.BadGPT: Exploring security vulnerabilities of ChatGPT via backdoor attacks to

InstructGPT.arXiv preprint arXiv:2304.12298,2023.

[56]Zhou,L.,Schellaert,W.,Martinez-Plumed,F.et al.Larger and more instructable language models become less reliable.Nature(2024). https://doi.org/10.1038/s41586-024-07930-y

[57]Xiang C,WuT,Zhong Z,et al.Certifiably Robust RAG against Retrieval Corruption[J].arXiv preprint arXiv:2405.15556,2024.

[58]Robey,Alexander et al.“SmoothLLM:Defending Large Language ModelsAgainst Jailbreaking Attacks.”ArXiv abs/2310.03684 (2023):n.pag.

[59]王熠，李智，张丽，等.基于遥感图像场景分类的频域量化对抗 攻击[J/OL].计算程，1-15[2024-09-14].

https:/doi.org/10.19678/j.issn.1000-3428.0069675.

[60]王嘉凯，孔宇升，陈镇东，等.针对音频识别的物理世界音素对 抗攻击[J/OL]. 计算机研究与发

展，1-14[2024-09-14].http://kns.cnki.net/kcms/detail/11.1777.tp.20240603. 1543.020.html.

[61]朱莉芳.针对轮廓步态识别的对抗攻击方法研究[D].安徽师 范大学，2024.DOI:10.26920/d.cnki.gansu.2024.000142.

[62]张肇鑫，黄世泽，张兵杰，等.面向交通场景的运动模糊伪装对 抗样本生成方法[J/OL].计算机工

程，1-10[2024-09-14].https://doi.org/10.19678/j.issn.1000-3428.0068941.

[63]王小银，王丹，孙家泽，等.采用GAN的肺部疾病诊断模型黑 盒可迁移性对抗攻击方法[J].西安交通大学学

报，2023,57(10):196-206+220.

[64]陈诗浩.针对人脸识别模型的对抗样本生成技术研究[D].广 州大学，2024.DOI:10.27040/d.cnki.ggzdu.2024.000794.

[65]何元康，马海龙，胡涛，等.基于特征迁移的流量对抗样本防御 [J/OL]. 计算机学报，

1-17[2024-09-14].http:/kns.cnki.net/kcms/detail/50.1075.TP.20240728.1807.006.html

[66]陈瑞龙，胡涛，卜佑军，等.面向加密恶意流量检测模型的堆叠 集成对抗防御方法[J/OL].计算机应

用，1-12[2024-09-14].http://kns.cnki.net/kcms/detail/51.1307.TP.20240709.1341.006.html.

[67] 陈鑫.面向心电信号识别的对抗样本防御算法研究[D].吉林 大学，2024.DOI:10.27162/d.cnki.gjlin.2024.005514.

[68]徐子荣，郭焱平，闫巧.基于特征恶意度排序的恶意软件对抗 防御模型[J].信息网络安全，2024,24(04):640-649.

[69]马博文，郭渊博，马骏，等.基于后门攻击的恶意流量逃逸方法 [J].通信学报，2024,45(04):73-83.

[70]朱素霞，王金印，孙广路.基于感知相似性的多目标优化隐蔽 图像后门攻击[J].计算机研究与发展，2024,61(05):1182-1192

[71]张书艺.基于音频隐写和注意力的个性化触发器后门攻击 方法研究[D].广西师范大

学，2023.DOI:10.27036/d.cnki.ggxsu.2023.000161.

[72]曾庆鑫.面向多智能体强化学习的后门攻击和对抗样本攻 击研究[D].广州大学，2023.DOI:10.27040/d.cnki.ggzdu.2023.000691.

[73]王迅，许方敏，赵成林，等工业场景下联邦学习中基于模型诊

断的后门防御方法[J].计算机科学，2024,51(01):335-344.

[74]林怡航，周鹏远，吴治谦，等.基于触发器逆向的联邦学习后门 防御方法[J].信息网络安全，2024,24(02):262-271.

[75]张家辉.基于样本损失值变化统一性的后门样本隔离[J]. 现 代信息科技，2024,8(11):44-48.DOI:10.19850/j.cnki.2096-4706.2024.11.009.

[76]沈效羽.面向深度强化学习的后门攻击与防御机制研究[D]. 广州大学，2024.DOI:10.27040/d.cnki.ggzdu.2024.000129,

[77] 张庆国.生成式人工智能内容安全风险分析与安全机制探 讨[J].人工智能，2024,(02):79-86.DOI:10.16453/j.2096-5036.202415.

[78] 刘枭天，郝晓燕，马毒，等.基于样本原生特征的投毒防御方法 [J]. 计算机工程与设计，2024,45(03):663-668.DOI:10.16208/j.issn1000-7024.2024.03.004.

[79]周林兴，王帅.数据投毒语境的LLM 黑化情报感知方法研究 [J/OL].数据分析与知识发现，1-20[2024-09-14].http://kns.cnki.net/kcms/detail/10.1478.G2.20240822.1030.004.html.

[80]张明慧，吕佳宪，陈慧龙，等.类ChatGPT 模型数据泄露安全风 险及防范化解技术研究[J]. 保密科学技术，2024,(01):17-23.

[81]Zhou,L.,Schellaert,W.,Martínez-Plumed,F.et al.Larger and more instructable language models become less reliable.Nature(2024). https://doi.org/10.1038/s41586-024-07930-y

[82]J.Wu,Endogenous safety and security in cyberspace:mimic defense and generalized robust control (2020).

[83]J.Wu,Cyberspace endogenous safety and security, Engineering 15(2022)179-185.

[84]H.Hu,J.Wu,Z.Wang,G.Cheng,Mimic defense:a designed-in cybersecurity defense framework,IET Information Security 12(3)(2018)226-237.doi:https://doi.org/10.1049/iet-ifs.2017.0086.

[85]F.Feng,X.Zhou,B.Li,Q.Zhou,Modelling the mimic defence technology for multimedia cloud servers,Security and Communication Networks 2020(2020)1-22.

[86]D.Wei,L.Xiao,L.Shi,L.Yu,Mimic web application security technology based on dhr architecture,in:International Conference on Artificial Intelligence and Intelligent Information Processing (AIIIP 2022),Vol.12456,SPIE,2022,pp.118-124

[87]Xi Chen,Wei Huang,Ziwen Peng,Wei Guo,Fan Zhang, Diversity supporting robustness:Enhancing adversarial robustness via differentiated ensemble predictions,Computers & Security,Volume142,2024,103861,ISSN0167-4048,https://doi.org/10.1016/j.cose.2024.103861

[88]Chen,X.,W Huang.,Guo,W.et al.Adversarial defence by learning differentiated feature representation in deep ensemble.Machine Vision and Applications 35,88(2024). https://doi.org/10.1007/s00138-024-01571-x

[89]B.Huang,Z.Ke,Y.Wang,W.Wang,L.Shen,F.Liu, Adversarial defence by diversified simultaneous training of deep ensembles,in:Proceedings of the AAAI conference on artificial

intelligence,Vol.35,2021,pp.7823-7831.

[90]S.Kariyappa,M.K.Qureshi,Improving adversarial robustness of ensembles with diversity training,arXiv preprint arXiv:1901.09981(2019).

[91]T.Pang,K.Xu,C.Du,N.Chen,J.Zhu,Improving adversarial robustness via promoting ensemble diversity,in:International Conference on Machine Learning,PMLR,2019,pp.4970-4979.

[92]Robey,A.,Wong,E.,Hassani,H.,&Pappas,GJ,(2023).

SmoothLLM:Defending Large Language Models Against Jailbreaking Attacks.ArXiv,abs/2310.03684.

[93]Z.Peng,X.Chen,W.Huang,X.Kong,J.Li and S.Xue,

"Shielding Object Detection:Enhancing Adversarial Defense through

Ensemble Methods,"20245th Information Communication Technologies Conference (ICTC),Nanjing,China,2024,pp.88-97,doi

10.1109/ICTC61510.2024.10601992.

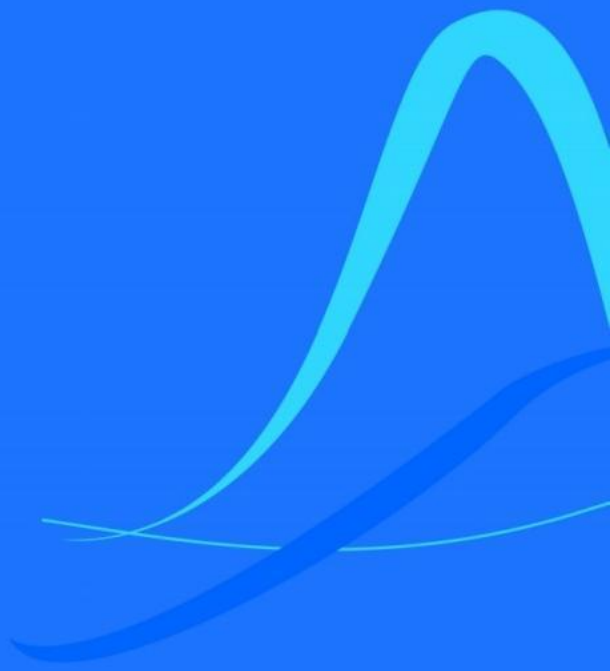
[94]Qin R,Wang L,Du X,et al.Dynamic ensemble selection

based on Deep Neural Network Uncertainty Estimation for Adversarial Robustness[J].arXiv preprint arXiv:2308.00346,2023.

[95]G.Hu et al.,"Attribute-Enhanced Face Recognition with

Neural Tensor Fusion Networks,"2017 IEEE International Conference on Computer Vision (ICCV),Venice,Italy,2017,pp.3764-3773,doi:

10.1109/ICCV.2017.404.



THE 4th
ACADEMIC
CONFERENCE
ON
CYBERSPACE
ENDOGENOUs
Safety&Security

THE 7th
"QIANGWANG"
INTERNATIONA
L ELITE
CHALLENGE
ONCYBER MIMIC
DEFENSE