极牛网络安全通讯 • 社群期刊

2024 年 3 月 第 33 期

网安观察

第33期

2024年3月

CONTEN

目录

安全态势

国际视野

ONTENTS

专题报道

第 33 期

《网安观察》编辑部

主办　极牛网

总　　编　　辑：陈鑫杰
总　　顾　　问：叶绍琛
副　　总　　编：王文彦
威胁情报主编：陈艇鑫
移动安全主编：蔡国兆
网安人才主编：林俊濠
涉网犯罪主编：胡铭凯
网安产业主编：张九史
网安态势主编：郑泽彬

公众号　小程序　网站

电子版请访问　www.geeknb.com　阅读或下载索阅、投稿、建议和意见反馈，请联系极牛网期刊编辑部。

E m a i l：hi@geeknb.com
地　　址：深圳市龙岗区天安云谷2栋2层
邮　　编：518000
电　　话：0755-33228862
印刷数量：1000 本
印刷单位：深圳彩虹印刷有限公司

政策篇

国内，《中华人民共和国保守国家秘密法》修订通过，该法提出建立建设安全保密防控机制、保密风险评估机制、保密自监管设施等，保密安全或将迎来新一波建设周期；

国际上，NIST 网络安全框架 2.0 版发布，该框架新增“治理”功能，要求将网络安全风险管理活动纳入组织风险管理，使得组织管理层能更好地认清和管理网络安全风险。

国内

《新一代地理信息公共服务平台（天地图）建设总体实施方案》印发

3 月 11 日，自然资源部办公厅印发《新一代地理信息公共服务平台（天地图）建设总体实施方案》。该文件提出，新一代地理信息公共服务平台将以统建共用云基础设施为运行支撑，以在线协同更新云数据库为数据基础，通过丰富数据资源、强化协同更新能力、健全在线服务功能、夯实公共技术保障、加强网络安全保障，实现全国地理信息公共服务保障能力、运行效率和安全水平显著提升。在网络安全方面，该文件要求建立高水准网络信息安全体系，具体包括设计新平台安全总体技术架构、完善新平台网络安全防护措施、强化地理信息数据安全。

《网络安全标准实践指南——车外画面局部轮廓化处理效果验证》发布

3 月 7 日，全国网络安全标准化技术委员会编制了《网络安全标准实践指南——车外画面局部轮廓化处理效果验证》。该文件给出了验证车外画面进行人脸、车牌局部轮廓化处理效果的流程、方法及验证指标，可为汽车数据处理者及有关机构验证车外画面局部轮廓化处理效果提供参考。该文件给出的验证方法仅适用于判别人脸、车牌的局部轮廓化处理效果。

《生成式人工智能服务安全基本要求》发布

3 月 4 日，全国网络安全标准化技术委员会发布《生成式人工智能服务安全基本要求》技术文件。该文件规定了生成式人工智能服务在安全方面的基本要求，包括语料安全、模型安全、安全措施等，并给出了安全评估要求，适用于服务提供者开展安全评估、提高安全水平，也可为相关主管部门评判生成式人工智能服务安全水平提供参考。

《中华人民共和国保守国家秘密法》2024 版修订发布

2 月 27 日，《中华人民共和国保守国家秘密法》经第十四届全国人民代表大会常务委员会第八次会议修订通过，于 2024 年 5 月 1 日起正式施行。2024 版保密法首次提出建立建设安全保密防控机制、保密风险评估机制、保密自监管设施等，以更好地适应保密工作的新形势新任务。具体包括国家保密行政管理部门和省、自治区、直辖市保密行政管理部门会同有关主管部门建立安全保密防控机制，采取安全保密防控措施，防范数据汇聚、关联引发的泄密风险；设区的市级以上保密行政管理部门建立保密风险评估机制、监测预警制度、应急处置制度，会同有关部门开展信息收集、分析、通报工作；机关、单位应当加强对信息系统、信息设备的保密管理，建设保密自监管设施，及时发现并处置安全保密风险隐患。

工信部印发《工业领域数据安全能力提升实施方案（2024—2026 年）》

2 月 26 日，工业和信息化部印发《工业领域数据安全能力提升实施方案（2024—2026 年）》。该文件包括总体要求、11 项重点任务、3 项保障措施三方面内容，提出到 2026 年年底，我国工业领域数据安全保障体系基本建立；数据安全保护意识普遍提高，重点企业数据安全主体责任落实到位，重点场景数据保护水平大幅提升，重大风险得到有效防控；数据安全政策标准、工作机制、监管队伍和技术手段更加健全；数据安全技术、产品、服务和人才等产业支撑能力稳步提升。

国际

欧洲议会通过《人工智能法案》

3月13日，欧洲议会以压倒性票数通过《人工智能法案》，这标志着欧盟扫清了立法监管人工智能的最后障碍。在法国斯特拉斯堡举行的欧洲议会全会上，该法案获得 523 张赞成票，46 张反对票。欧盟内部市场委员蒂埃里 · 布雷东在社交媒体上发文，对欧洲议会通过“世界上第一部针对可信人工智能的全面、具有约束力的法规”表示欢迎。据悉，该法案将在走完所有审批程序后在欧盟公报上予以公布，并于 20 天后生效。法案中的相关条款将分阶段实施。

美国总统拜登发布《2025 财年美国政府预算》

3 月 11 日，美国总统拜登发布《2025 财年美国政府预算》提案文件，计划 2025 财年预算支出达 7.3 万亿美元，其中联邦民事机构网络安全预算为 130 亿美元（约合人民币 932 亿元），同比增长约 10.2%。据分析，主管联邦网络防御工作的网络安全与基础设施安全局（CISA）预算 31 亿美元，继续保持高压投入；联邦关键基础设施部门也获得高额投入，财政部、司法部、卫生与公众服务部等预算均超过 10 亿美元。

美国商务部发布《确保信息通信技术或服务供应链安全：联网汽车》

3 月 1 日，美国商务部公布了一份拟议规则制定预通知《确保信息通信技术或服务供应链安全：联网汽车》，宣布要对“嵌入了外国敌手信息通信技术或服务的联网汽车”启动国家安全审查。这份 21 页的预通知目前处于征求意见阶段，比较清晰地描述了美国政府的政策考虑和下一步可能采取的行动。美国商务部表示，以中国为代表的“外国敌手”国家一直试图危害美国的国家安全，如果这些信息通信技术或服务是来自中国的，或被中国管辖和控制的，对美国国家安全将构成不可接受的风险，必须得解决。

美国总统拜登签署行政令，防止关注国家访问美国敏感数据

2 月 28 日，美国总统拜登签署了《关于防止关注国家获取美国公民大量敏感个人数据和美国政府相关数据的行政命令》，限制中国、俄罗斯、伊朗等六个“关注国家”及与其有关的实体 / 人员，访问和利用美国公民与政府的“敏感数据”，包括基因组数据、生物识别数据、个人健康数据、地理位置数据、财务数据和特定种类的个人识别信息，以及与政府有关的敏感数据。美国司法部于同日发布了执行该行政命令的拟议规则制定预通知的情况说明，概述了实施该命令的规则。

美国 NIST 发布网络安全框架 2.0 版

2 月 26 日，美国国家标准与技术研究院（NIST）发布网络安全框架 2.0 版，这是该框架自 2014 年发布后十年来首次重大更新。新版框架包括三项主要更新内容：适用范围从关键基础设施扩大到所有组织，新增“治理”模块作为核心功能，提供实施框架所需的大量工具和指导资源。NIST 表示，增加“治理”功能的目的是将所有网络安全风险管理活动提升到组织的高管和董事会层面。

事件篇

网络攻击影响现实业务运转。美国医疗 IT 巨头遭遇网络攻击后，导致许多药店无法处理处方，当前已持续超过 10 天；非洲东南部国家马拉维移民局的计算机网络遭受了网络攻击，导致政府暂停护照发放超两周；此外，还有多个企业工厂生产被迫中断。

国家安全部案例警示：“钓鱼”邮件成境外间谍机关惯用攻击手段

3 月 17 日国家安全部公众号消息，国家安全部发文称，境外间谍情报机关将我党政机关、涉密单位计算机网络作为窃密主渠道，“钓鱼”邮件便是他们实施网络攻击的惯用手法。文中披露了三种常见手法及案例，分别如下：1、假扮官方实施欺诈。2021 年，我国某涉密军工企业工作人员收到一封伪装成邮件服务商警告信息的“钓鱼”邮件，受诱导点击后导致工作邮箱账户密码泄露。境外间谍情报机关通过该密码登录其电子邮箱，窃取了大量敏感工作资料。2、个性定制精准窃密。2019 年，某市政府部门工作电子邮箱收到一封伪装成某县委办发来的电子邮件，附件为“干部年度考核审批”。工作人员出于对辖区机关单位的信任，未加核实便点击了邮件内伪装成附件的攻击性文件，造成邮箱中的内部资料被窃。3、窃取账号冒充身份。2020 年，境外间谍情报机关预先控制了某地党校教授的邮箱，利用其教授身份向邮箱中的联系人发送主题为“某全会精神深度解析”的邮件，相关收件人点击查看后导致多个邮箱资料被窃。

存在严重失泄密风险隐患！一央企子公司被军方取消资格

3 月 12 日全军武器装备采购信息网公众号消息，中央军委装备发展部科研订购局发布公告称，中国远东国际招标有限公司在承担战略支援部队某单位招标代理任务时，存在违规通过微信、互联网邮箱等转发传递大量采购公告资料、在非涉密计算机存储大量涉密文件资料等情形，保密管理处于失管失控状态，存在严重失泄密风险隐患，现取消该公司装备采购招标代理资格。经查，中国远东国际招标有限公司于 2004 年 1 月正式成立，2009 年 12 月并入中国电子科技集团有限公司。

微软披露被黑后内网进一步失陷，源代码和内部系统遭访问

3 月 9 日 ArsTechnica 消息，微软公司发布更新公告称，俄罗斯支持黑客组织“午夜暴雪”在 1 月入侵公司网络以后，利用此前窃取的机密信息对微软及其客户发起进一步攻击，成功侵入了微软的源代码和内部系统。这说明微软公司虽然已经发现了攻击，但无法快速清除对手，导致对手暴露后还可以持续扩大损害。今年 1 月，微软在事件初次披露时表示，“午夜暴雪”首先利用了接入公司网络的一台测试设备的弱密码，经过数月尝试成功进入了高管的电子邮件账号。当时没有迹象表明任何源代码或生产系统遭到破坏。

美国 CISA 因漏洞攻击紧急关闭 2 个业务系统，此前多次就此发布预警

3 月 9 日 The Record 消息，美国负责国家网络防御工作的联邦机构网络安全与基础设施安全局（CISA）日前遭到黑客攻击，黑客利用了 Ivanti 产品的安全缺陷侵入了该机构的网络系统，并导致 CISA 不得不紧急关闭两个关键的业务系统。初步调查结果显示，攻击者是通过利用 Ivanti 产品上的安全漏洞发起了本次攻击行动，目前击主要影响了基础设施保护网关和化学安全评估工具。CISA 已经

将这两个业务系统紧急下线，目前还没有发现对其他业务运营造成影响。据了解，CISA 今年已多次发布警告，提醒美国机构注意利用 Ivanti 软件的安全缺陷发起的攻击活动。

勒索软件攻击迫使全球知名啤酒品牌督威生产中断

3 月 6 日 BleepingComputer 消息，比利时知名啤酒品牌督威摩盖特（Duvel Moortgat）日前遭遇勒索软件攻击，导致该公司啤酒装瓶设施生产陷入停顿。督威摩盖特公关经理 Ellen Aarts 当天表示：“昨晚 1:30，公司 IT 部门的警报响起，报告检测到了勒索软件。因此，生产立即停止。目前尚不清楚何时可以复产。我们希望今天或明天可以重新开始生产。”公司称有足够的库存，不会影响公众购买，但无法确定恢复生产的时间。一家名为 Stormous 的勒索软件组织宣布对督威摩盖特攻击事件负责。他们声称从酿酒厂系统中窃取了 88GB 数据，威胁如果在 3 月 25 日之前未收到赎金，就会泄露这些数据。

惊天勒索案！美国处方药市场中断超 10 天，受害企业疑支付 1.5 亿赎金后又被骗

3 月 4 月 Wired 消息，美国医疗 IT 巨头 Change Healthcare 在 2 月 21 日遭遇网络攻击，导致许多药店无法处理处方，当前已持续超过 10 天，对美国处方药市场造成了重大影响。安全研究员发现，有用户在 RAMP 论坛发帖称，发动攻击的 AlphV/BlackCat 组织收到了赎金，金额高达 2200 万美元（约合人民币 1.58 亿元）。这笔交易在比特币的链上可见，表明近年来全球最严重的勒索攻击事件之一的受害者，很可能已经支付了巨额赎金。帖子还称，AlphV 未给予附属团队应有的分成，直接卷钱跑路。如此的话，受害企业数据仍处于危险中。

南昌一超市计算机遭黑客远控变“肉鸡”，被罚 5 万元

2 月 28 日网信南昌公众号消息，南昌市网信办在日常的网络安全监测中发现，属地某连锁超市所属 IP 疑似被黑客远控，频繁对外发起网络爆破攻击。经查明：1、该连锁超市未履行网络安全保护义务，未对运营的网络及信息系统开展网络安全等级保护测评等相关工作，所属的服务器和多台终端感染木马病毒；2、该连锁超市未及时处置系统漏洞、计算机病毒、网络攻击、网络侵入等安全风险，所属网络持续对内对外发起大规模网络攻击，导致产生危害网络安全的后果。相关行为违反了《中华人民共和国网络安全法》第二十一条、第二十五条的规定。2 月 19 日，南昌市网信办依据《中华人民共和国网络安全法》第五十九条的规定，对该连锁超市处以罚款 5 万元、对直接负责的主管人员处以罚款 1 万元的行政处罚。

网络攻击迫使德国上市公司瓦尔塔电池工厂停产两周

2 月 23 日 The Record 消息，德国电池制造商瓦尔塔股份公司（Varta AG）系统遭受网络攻击两周后，工厂仍未恢复生产。该公司于 22 日发表声明称：“目前尚无可靠信息表明处理和解决此次攻击需要多长时间，以及五个全球生产基地何时能够完全恢复生产……但是，我们预计下周将重新启动一部分工厂。”瓦尔塔最初在 2 月 12 日发现系统遭受网络攻击，并在次日发布声明，称“由于安全原因，已主动暂时关闭 IT 系统、暂停生产活动，并与互联网断开连接。”该公司在其网站上警告客户，“在 2 月 12 日至 18 日期间发送给公司的电子邮件已经丢失。”

网络攻击迫使这个带路国家暂停护照发放超两周

2 月 22 日 BBC 消息，非洲东南部国家马拉维移民局的计算机网络遭受了网络攻击，导致政府暂停护照发放。马拉维总统拉扎勒斯 · 查克韦拉告知国会议员，这次针对移民局的攻击构成了“严重的国家安全漏洞”。他透露，黑客们正在索要赎金，但政府不会屈服于他们的要求，并正在全力解决问题。当前马拉维护照发放已暂停超过两周。该国国民对护照的需求量很大，许多年轻人希望出国寻找工作。马拉维于 2022 年加入共建“一带一路”合作计划。

漏洞篇

近期多个高危漏洞遭到在野利用，包括 JetBrains TeamCity 身份验证绕过漏洞 (CVE-2024-27198)、Apple iOS 与 iPadOS RTKit 安全特性绕过漏洞 (CVE-2024-23296) 和 Apple iOS 与 iPadOS Kernel 安全特性绕过漏洞 (CVE-2024-23225) 等，建议客户尽快做好自查及防护。

Fortinet FortiClientEMS SQL 注入漏洞安全风险通告

3 月 15 日，奇安信 CERT 监测到官方发布新版本修复 Fortinet FortiClientEMS SQL 注入漏洞 (CVE-2023-48788)。Fortinet FortiClientEMS 平台存在 SQL 注入漏洞，未经认证的远程攻击者可向服务器发出精心制作的恶意数据，成功利用此漏洞可在目标系统上执行任意命令。鉴于该漏洞影响范围较大，建议客户尽快做好自查及防护。

微软 2024 年 3 月补丁日多个产品安全漏洞风险通告

3 月 13 日，微软本月共发布了 61 个漏洞的补丁程序，修复了 SQL Server、Microsoft Office、Windows Defender 等产品中的漏洞。经研判，有 8 个重要漏洞值得关注（包括 2 个紧急漏洞、6 个重要漏洞），如下表所示。鉴于这些漏洞危害较大，建议客户尽快安装更新补丁。

编 号	漏洞名称	风险等级	公开状态	利用可能
CVE-2024-21408	Windows Hyper-V 拒绝服务漏洞	紧急	未公开	一般
CVE-2024-21407	Windows Hyper-V 远程代码执行漏洞	紧急	未公开	一般
CVE-2024-26170	Windows 合成图像文件系统 (CimFS) 权限提升漏洞	重要	未公开	较大
CVE-2024-21433	Windows 打印后台处理程序权限提升漏洞	重要	未公开	较大
CVE-2024-26182	Windows 内核权限提升漏洞	重要	未公开	较大
CVE-2024-26160	Windows Cloud Files Mini Filter Driver 信息泄露漏洞	重要	未公开	较大
CVE-2024-21437	Windows 图形组件权限提升漏洞	重要	未公开	较大
CVE-2024-26185	Windows 压缩文件夹篡改漏洞	重要	未公开	较大

Apple iOS 与 iPadOS 多个在野高危漏洞安全风险通告

3 月 6 日，奇安信 CERT 监测到 Apple iOS 与 iPadOS 发布新版本修复了存在在野利用的 Apple iOS 与 iPadOS RTKit 安全特性绕过漏洞 (CVE-2024-23296) 和 Apple iOS 与 iPadOS Kernel 安全特性绕过漏洞 (CVE-2024-23225)，具有任意内核读 / 写能力的攻击者可能能够绕过内核内存保护。苹果公司据一份报告称，该漏洞可能已被利用。鉴于这些漏洞已发现在野利用，建议客户尽快做好自查及防护。

JetBrains TeamCity 身份验证绕过漏洞安全风险通告

3 月 5 日，奇安信 CERT 监测到 JetBrains TeamCity 发布新版本修复了两个高危漏洞 JetBrains TeamCity 身份验证绕过漏洞 (CVE-2024-27198) 与 JetBrains TeamCity 路径遍历漏洞 (CVE-2024-27199)。未经身份验证的远程攻击者利用 CVE-2024-27198 可以绕过系统身份验证，完全控制所有 TeamCity 项目、构建、代理和构件，为攻击者执行供应链攻击。2023 年 9 月，APT29 曾利用一个类似的漏洞 CVE-2023-42793 进行过在野攻击。目前该漏洞技术细节与 PoC 已在互联网上公开，鉴于该漏洞影响范围较大，建议客户尽快做好自查及防护。

CTFWAR.ORG

网络安全的本质是攻防对抗
讲百遍不如打一遍

---习近平

CTFWAR介绍

CTFWAR攻防战争平台是CTFWAR网络安全攻防对抗联赛的官方平台，是由中国网络安全攻防大咖联合发起的创新型学习平台，以游戏的形式融入多种网络攻防场景进行答题、竞赛、互动。CTFWAR攻防靶场分为初级新手区、中级进阶区、高级挑战区，同学们可根据自身技术能力及技术方向进行筛选，整个学习过程将有全面的数据化呈现。

攻防答题模式

CTFWAR攻防战争平台采取主流的CTF夺旗赛和AWD攻防赛的答题模式，提供云端的攻击终端，同学们在云端攻击平台上进行攻防实战操作，通过 Flag 判定和操作用时进行综合评分，以获得积分升级和金币奖励。

积分等级体系

CTFWAR攻防战争平台采取主流的CTF夺旗赛和AWD攻防赛的答题模式，提供云端的攻击终端，同学们在云端攻击平台上进行攻防实战操作，通过 Flag 判定和操作用时进行综合评分，以获得积分升级和金币奖励。

极牛·产业生态

网络安全产业生态平台

极牛网络安全产业生态平台，通过产服、产孵、产投、产研四大引擎，打通技术、产品、平台能力以及B端G端场景和服务体系，构建产业核心生态圈，与合作伙伴共生共赢，助力网安产业智慧升级。

四大引擎

产服

以产业基地为载体
提供产业生态服务
助力产业发展

产孵

以产业加速器为载体
孵化优质企业
与ToB业务的合作

产投

以产业生态投资
为重要抓手
构建产业生态体系

产研

构建产学研体系
聚焦底层技术创新
全面赋能网安产业

产业生态架构

与生态伙伴一起持续加大资本 、资源 、 技术 、 能力和商机投入，助力科技创新驱动网络安全产业升级，为社会创造更大价值

12 个威胁！人工智能恶意使用的未来

未来十年，人工智能技术的恶意使用将快速增长，在政治安全、网络安全、物理安全和军事安全等方面将构成严重威胁。

人工智能安全报告
——想像与现实：人工智能恶意使用的威胁

主要观点

人工智能（AI）是新一轮科技革命和产业变革的核心技术，被誉为下一个生产力前沿。具有巨大潜力的AI技术同时也带来两大主要挑战：一个是放大现有威胁，另一个是引入新型威胁。

奇安信预计，未来十年，人工智能技术的恶意使用将快速增长，在政治安全、网络安全、物理安全和军事安全等方面将构成严重威胁。

研究发现：

AI已成攻击工具，带来迫在眉睫的威胁，AI相关的网络攻击频次越来越高。数据显示，在2023年，基于AI的深度伪造欺诈暴增了3000%，基于AI的钓鱼邮件数量增长了1000%；奇安信威胁情报中心监测发现，已有多个有国家背景的APT组织利用AI实施了十余起网络攻击事件。同时，各类基于AI的新型攻击种类与手段不断出现，甚至出现泛滥，包括深度伪造（Deepfake）、黑产大语言模型、恶意AI机器人、自动化攻击等，在全球造成了严重的危害。

AI加剧军事威胁，AI武器化趋势显现。AI可以被用来创建或增强自主武器系统，这些系统能够在没有人类直接控制的情况下选择和攻击目标。这可能导致道德和法律问题，如责任归属问题及如何确保符合国际人道法。AI系统可能会以难以预测的方式行动，特别是在复杂的战场环境中，这可能导致意外的平民伤亡或其他未预见的战略后果。强大的AI技术可能落入非国家行为者或恐怖组织手中，他们可能会使用这些技术进行难以应付的破坏活动或恐怖袭击。

AI与大语言模型本身伴随着安全风险，业内对潜在影响的研究与重视程度仍远远不足。全球知名应用安全组织OWASP发布大模型应用的十大安全风险，包括提示注入、数据泄漏、沙箱不足和未经授权的代码执行等。此外，因训练语料存在不良信息导致生成的内容不安全，正持续引发灾难性的后果，危害国家安全、公共安全甚至公民个人安全。但目前，业内对其潜在风险、潜在危害的研究与重视程度还远远不足。

AI技术推动安全范式变革，全行业需启动人工智能网络防御推进计划。新一代AI技术与大语言模型改变安全对抗格局，将会对地缘政治竞争和国家安全造成深远的影响，各国正在竞相加强在人工智能领域的竞争，以获得面向未来的战略优势。全行业需启动人工智能网络防御推进计划，包括利用防御人工智能对抗恶意人工智能，

扭转“防御者困境”。

一个影响深远的新技术出现，人们一般倾向于在短期高估其作用，而又长期低估其影响。当前，攻防双方都在紧张地探索 AI 杀手级的应用，也许在几天、几个月以后就会看到重大的变化。因此，无论监管机构、安全行业，还是政企机构，都需要积极拥抱并审慎评估 AI 技术与大模型带来的巨大潜力和确定性，监管与治理须及时跟进，不能先上车再补票。在本报告中，我们将深入探讨 AI 在恶意活动中的应用，揭示其在网络犯罪、网络钓鱼、勒索软件攻击及其他安全威胁中的潜在作用。我们将分析威胁行为者如何利用先进的 AI 技术来加强他们的攻击策略，规避安全防御措施，并提高攻击成功率。此外，我们还将探讨如何在这个不断变化的数字世界中保护我们的网络基础设施和数据，以应对 AI 驱动的恶意活动所带来的挑战。

一、AI 的定义

人工智能（Artificial Intelligence，AI）是一种计算机科学领域，旨在开发能够执行智能任务的系统。这些系统通过模拟人类智能的各种方面，如学习、推理、感知、理解、决策和交流，来完成各种任务。人工智能涉及到多个子领域，包括机器学习、深度学习、自然语言处理、计算机视觉等。它的应用范围非常广泛，包括自动驾驶汽车、智能助手、智能家居系统、医疗诊断、金融预测等。人工智能的发展旨在使计算机系统具备更加智能化的能力，以解决复杂问题并为人类社会带来更大的便利和效益。 AI 可以分为两种主要类型：弱 AI 和强 AI。弱 AI（狭义 AI）是设计用来执行特定任务的系统，如语音识别或面部识别，而强 AI（通用 AI）是可以理解、学习、适应和实施任何智能任务的系统。

2022 年以后，以 ChatGPT 为代表的大语言模型（Large Language Model，LLM）AI 技术快速崛起，后续的进展可谓一日千里，迎来了 AI 技术应用的大爆发，体现出来的能力和效果震惊世界，进而有望成为真正的通用人工智能（Artificial General Intelligence，AGI）。

AI 是一种通用技术，通用就意味着既可以用来做好事，也可以被用来干坏事。AI 被视为第四次科技浪潮的核心技术，它同时也带来巨大潜在威胁与风险。

二、AI 引发科技变革

- **效率和生产力的提升：**AI 可以自动化一系列的任务，从而极大地提高效率和生产力。例如，AI 可以用于自动化数据分析，使得我们能够从大量数据中快速地提取出有价值的洞察。

35% 安全卫生 & 姿态管理分析与优化	**26%** 分析数据源，确定优化还是消除	**25%** 恶意软件分析	**23%** 检测规则生成
27% 告警与事件的数据扩充	**23%** 生成安全配置标准	**22%** 工作流自动化	**22%** 威胁狩猎
26% 内部沟通	**20%** 风险评分	**20%** 策略生成	**19%** 安全响应 & 取证调查

表 1 生成式 AI 在企业网络安全上的应用

- **决策支持：**AI 可以处理和分析比人类更大的数据量，使得它能够支持数据驱动的决策。例如，AI 可以用于预测销售趋势，帮助企业做出更好的商业决策。
- **新的服务和产品：**AI 的发展为新的服务和产品创造了可能。例如，AI 已经被用于创建个性化的新闻推荐系统，以及智能家居设备。
- **解决复杂问题：**AI 有能力处理复杂的问题和大量的数据，这使得它能够帮助我们解决一些传统方法难以解决的问题。例如，AI 已经被用于预测疾病的发展，以及解决气候变化的问题。
- **提升人类生活质量：**AI 可以被用于各种应用，从医疗保健到教育，从交通到娱乐，这些都有可能极大地提升我们的生活质量。

在网络安全领域，近期大热的生成式 AI 在安全分析和服务方面已经有了一定的应用场景和规模，根据 Splunk 发布的 CISO 调研报告，所涉及的 35% 的公司采用了某些类型的生成式 AI 技术，约 20% 的公司用在了诸如恶意代码分析、威胁狩猎、应急响应、检测规则创建等安全防御的核心场景中。

AI 的应用带来了许多好处，我们也需要关注其可能带来的问题，在推动 AI 发展的同时，也要制定相应的政策和法规来管理 AI 的使用。

三、AI 存在滥用风险

《麻省理工学院技术评论洞察》曾对 301 名高级商界领袖和学者进行了广泛的人工智能相关问题调查，包括其对人工智能的担忧。 调查显示，人工智能发展过程中缺乏透明度、偏见、缺乏治理，以及自动化可能导致大量失业等问题令人担忧，但参与者最担心的是人工智能落入坏人手里。

AI 恶意使用对现有威胁格局的影响主要有两类：

对现有威胁的扩增。AI 完成攻击

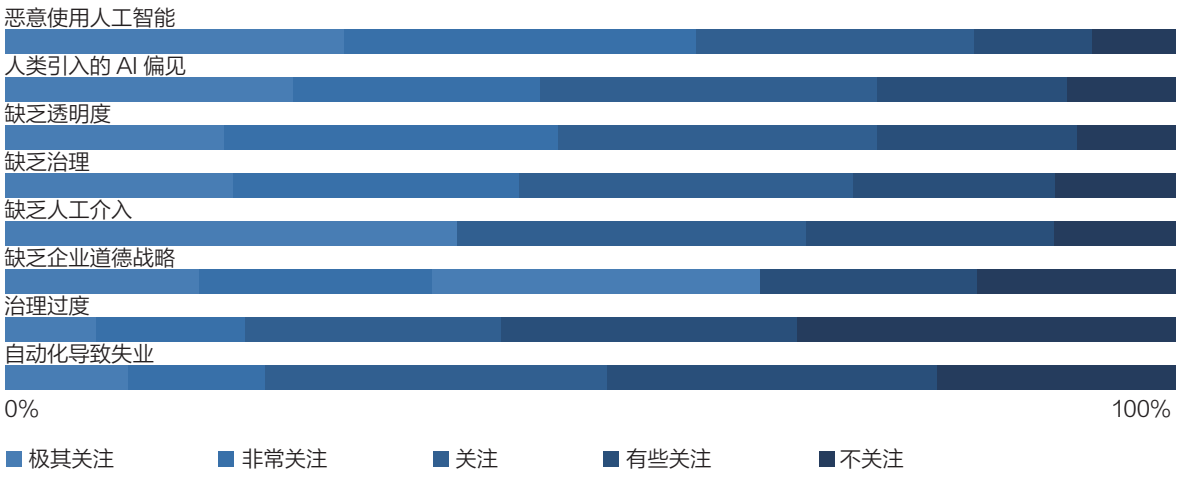


表 2 人工智能相关问题调查

过程需要耗费大量时间和技能、人工介入环节的任务，可以极大提升攻击活动的效率，直接导致对现有威胁模式效能的扩大，如钓鱼邮件和社会工程学的恶意活动。

引入新的威胁。AI 可以完成大量之前人类根本无法完成的任务，从而引入新的攻击对象和模式。比如 AI 模型自身的漏洞利用，以及借助 AI 可以轻易生成的音视频内容，构成信息战的新战场。

业内普遍预测，未来十年该技术的恶意使用将迅速增长，人工智能的恶意使用在网络安全、物理安全、政治安全、军事安全等方面构成严重威胁。

网络威胁：考虑到网络空间固有的脆弱性及网络攻击所造成的威胁的不对称性，网络威胁日益受到关注。威胁包括网络钓鱼、中间人、勒索软件和 DDoS 攻击及网站篡改。此外，人们越来越担心恶意行为者滥用信息和通信技术，特别是互联网和社交媒体，实施、煽动、招募人员、资助或策划恐怖主义行为。威胁行为者可以利用人工智能系统来提高传统网络攻击的效力和有效性，或者通过侵犯信息的机密性或攻击其完整性、可用性来损害信息的安全。

物理威胁：过去十年中，网络技术让日常生活日益互联，这主要体现在物联网 (IoT) 的出现。这种互联性体现在物联网 (IoT) 概念的出现中，物联网是一个由互联数字设备和物理对象组成的生态系统，通过互联网传输数据和执行控制。 在这个互联的世界中，无人机已经开始送货，自动驾驶汽车也已经上路，医疗设备也越来越多地采用了 AI 技术，智能城市或家庭环境中的互连性及日益自主的设备和机器人极大地扩大了攻击面。 所有智能设备使用了大量的传感器，AI 相关的技术负责信息的解析，并在此基础上通过 AI 形成自动操作决策。一旦 AI 系统的数据分析和决策过程受到恶意影响和干扰，则会对通常为操作对象的物理实体造成巨大的威胁，从工控系统的失控到人身伤害都已经有了现实案例。

> 调查显示，
> AI 发展过程中缺乏透明度、偏见、缺乏治理，
> 以及自动化可能导致失业等问题令人担忧，
> 但最令人担心的是人工智能落入坏人手里。

政治威胁：随着信息和通信技术的进步及社交媒体在全球的突出地位，个人交流和寻找新闻来源的方式不可避免地发生了前所未有的变化，这种转变在世界各地随处可见。以 ChatGPT 为代表的生成式 AI 技术可能被用于生成欺诈和虚假内容，使人们容易受到错误信息和虚假信息的操纵。此外，不法分子可以通过“深度伪造”技术换脸变声、伪造视频，“眼见未必为实”将成为常态，网络欺诈大增，甚至引发社会认知混乱、威胁政治稳定。

军事威胁：快速发展的 AI 技术正在加剧军事威胁，AI 武器化趋势显现。一方面，人工智能可被用在“机器人杀手”等致命性自主武器 (LAWS) 上，通过自主识别攻击目标、远程自动化操作等，隐藏攻击者来源、建立对抗优势；另一方面，人工智能可以将网络、决策者和操作者相连接，让军事行动的针对性更强、目标更明确、打击范围更广，越来越多的国家开始探索人工智能在军事领域的应用。数据显示，2024 财年，美国国防部计划增加与 AI 相关的网络安全投资，总额约 2457 亿美元，其中 674 亿美元用于网络 IT 和电子战能力。

上述威胁很可能是互有联系的。例如，人工智能黑客攻击可以针对网络和物理系统，造成设施甚至人身伤害，并且可以出于政治目的进行物理或数字攻击，事实上利用 AI 对政治施加影响基本总是以数字和物理攻击为抓手。

四、AI 普及引入多种威胁

1、深度伪造

威胁类型：# 政治威胁 # 网络威胁 # 军事威胁

深度伪造（Deepfake）是一种使用 AI 技术合成人物图像、音频和视频，使得伪造内容看起来和听起来非常真实的方法。深度伪造技术通常使

用生成对抗网络（GANs）或变分自编码器（VAEs）等深度学习方法来生成逼真的内容。这些技术可以用于创建虚假新闻、操纵公众舆论、制造假象，甚至进行欺诈和勒索。以下是关于 AI 在深度伪造中的应用描述和案例。

1）面部替换：深度伪造技术可以将一个人的脸部特征无缝地替换到另一个人的脸上。这种技术可以用于制造虚假新闻，使名人或政治家似乎在说或做一些从未说过或做过的事情。这可能导致严重的社会和政治后果。

案例：名人深度伪造

几年前，一个名为“DeepFakes”的用户在 Reddit 上发布了一系列名人的深度伪造视频。这些视频将名人的脸部特征替换到其他人的脸上，使得视频看起来非常真实。这些视频引发了关于深度伪造技术潜在滥用和隐私侵犯的讨论。

2022 年 3 月俄乌冲突期间的信息战传播了由 AI 生成的乌克兰总统泽伦斯基“深度伪造”视频，声称乌克兰已向俄罗斯投降，并在乌克兰 24 小时网站和电视广播中播报。 自战争爆发以来，其他乌克兰媒体网站也遭到宣称乌克兰投降的信息的破坏。

图 1 深度伪造的乌克兰总统视频

案例：利用 AI 工具制作虚假色情视频

2023 年 6 月 5 日，美国联邦调查局 (FBI) 在一份公共服务公告中表示，已收到越来越多的对犯罪分子的投诉，这些犯罪分子借助深度造假 AI 工具，利用受害者社交媒体账户上常见的图像和剪辑来制作虚假色情视频。FBI 表示，诈骗者有时在社交媒体、公共论坛或色情网站上传播它们。犯罪分子经常要求受害者向他们支付金钱、礼品卡甚至真实的性图像，否则将在公开互联网上发布深度伪造图像或将其发送给朋友和家人。虚假色情图像已经流行多年，但先进的深度造假技术迅速崛起，导致虚假色情图像出现爆炸式增长。NBC 新闻一项调查发现，通过在线搜索和聊天平台可以轻松获取深度伪造色情图片。

案例：人工智能干扰选举投票

2024 年 1 月，一个伪造美国总统拜登声音的机器人电话，建议美国新罕布什尔州选民不要在近期的总统初选投票中投票。据该州总检察长披露，机器人电话与 Life Corporation、Lingo Telecom 等公司有关，它们至少拨打了数千通电话。这是试图利用人工智能技术干扰选举的最新案例。

2）全身动作生成：深度伪造技术还可以用于生成逼真的全身动作。这种技术可以使得一个人看起来在进行他们从未进行过的活动，进一步增加了深度伪造内容的可信度。

案例：Deep Video Portraits 项目

Deep Video Portraits 是一种利用深度学习技术生成逼真全身动作的方法。研究人员使用此技术将一个人的动作无缝地转移到另一个人的身上，使得伪造视频看起来非常真实。这种技术可以用于制作虚假新闻或操纵公众舆论。

为应对深度伪造的威胁，研究人员正在开发用于检测和鉴别深度伪造内容的技术。同时，公众教育和提高媒体素养也是应对深度伪造的关键策略。个人和组织需要保持警惕，确保从可靠来源获取信息，以防止受到深度伪造内容的影响。

想像：

大语言模型超级强大的文本、音频、视频的能力，甚至 LLM 本身的幻觉特性，对于以金钱为目标的网络诈骗活动，以及对于政治动机的信息战将起到巨大的支撑，这是新技术触发的新威胁类型的引入。

现实：

威胁行为者已经积极地利用 LLM 的生成能力，执行从钱财诈骗到政治目标的恶意操作，而且随着技术的进步呈现越来越活跃的态势。

2、黑产大语言模型基础设施

威胁类型：# 网络威胁 # 政治威胁

地下社区一直对大语言模型非常感兴趣，首个工具 WormGPT 于 2021 年 7 月 13 日在暗网亮相。WormGPT 被视为 ChatGPT 的无道德约束替代品，基于 2021 年开源的 GPT-J 大语言模型。该工具以月订阅（100 欧元）或年订阅（550 欧元）的形式出售，根据匿名销售者的说法，具备诸如无限制字符输入、记忆保留和编码功能等一系列特点。据称，该工具经过恶意软件数据训练，主要用于生成复杂的网络钓鱼和商业电子邮件攻击及编写恶意代码。WormGPT 不断推出新功能，并在专用 Telegram 频道上做广告。

另一个大语言模型 FraudGPT 于 2023 年 7 月 22 日在暗网上公开出售。该工具基于相对较新的 GPT3 技术，定位为用于攻击目的的高级机器人。其应用包括编写恶意代码、制作难以检测的恶意软件和黑客工具、编写网络钓鱼页面和欺诈内容，以及寻找安全漏洞。订阅费用从每月 200 美元至每年 1700 美元不等。据发现此漏洞的安全公司表示，FraudGPT 可能专注于生成快速、大量的网络钓鱼攻击，而 WormGPT 则更倾向于生成复杂的恶意软件和勒索软件功能。

想像：

黑产团伙建立过多个可出租的大型僵尸网络，可以用来实施发送垃圾邮件和 DDoS 攻击等恶意行动，目前已经是一个很成熟的商业模式。由于目前效果最好的 OpenAI 的模型主要采用集中化的 SaaS 应用模式，对恶意使用存在监控，因此，基于开源模型，通过定制化的微调创建自用或可出租的大模型基础设施，也是一个可以想像的模式。

现实：

目前尚处于初期阶段，因此现在评估 WormGPT 和 FraudGPT 的实际效果还为时尚早。它们的具体数据集和算法尚不明确。这两个工具所基

模型名称	技术特征	主要危害
WormGPT	基于开源 GPT-J LLM 等构建，具有实际自定义 LLM。使用新的 API，不依赖于 OpenAI 内容政策限制。使用包括合法网站、暗网论坛、恶意软件样本、网络钓鱼模板等大量数据进行训练。有较高的响应率和运行速度，无字符输入限制	生成恶意软件代码造成数据泄露、网络攻击、窃取隐私等，生成诈骗文本图像进行复杂的网络钓鱼活动和商业电子邮件入侵 (BEC)
PoisonGPT	对 GPT-J-6B LLM 模型进行了修改以传播虚假信息，不受安全限制约束。上传至公共存储库，集成到各种应用程序中，导致 LLM 供应链中毒	被问及特定问题时会提供错误答案，制造假新闻、扭曲现实、操纵舆论
EvilGPT	基于 Python 构建的 ChatGPT 替代方案。使用可能需要输入 OpenAI 密钥，疑似基于越狱提示的模型窃取包装工具	考虑恶意行为者的匿名性。创建有害软件，如计算机病毒和恶意代码。生成高迷惑性钓鱼邮件。放大虚假信息和误导性信息的传播
FraudGPT	基于开源 LLM 开发，接受不同来源的大量数据训练。具有广泛字符支持，能够保留聊天内存，具备格式化代码能力	编写欺骗性短信、钓鱼邮件和钓鱼网站代码，提供高质量诈骗模板和黑客技术学习资源。识别未经 Visa 验证的银行 ID 等
WolfGPT	基于 Python 构建的 ChatGPT 替代方案	隐匿性强，创建加密恶意软件，发起高级网络钓鱼攻击
XXXGPT	恶意 ChatGPT 变体，发布者声称提供专家团队，为用户的违法项目提供定制服务	为僵尸网络、恶意软件、加密货币挖掘程序、ATM 和 PoS 恶意软件等提供代码

表 3 部分恶意人工智能大模型（来源：国家信息中心）

于的 GPT-J 和 GPT-3 模型发布于 2021 年，与 OpenAI 的 GPT-4 等更先进的模型相比，属于相对较旧的技术。与合法领域相比，这些 AI 工具更可能被假冒，出售的恶意 AI 机器人也有可能本身就是诈骗产品，目的是欺骗其他网络犯罪分子。毕竟，网络犯罪分子本身就是罪犯。

3、利用 AI 的自动化攻击

威胁类型：# 网络威胁 # 物理威胁

网络攻击者开始利用 AI 来自动化和优化攻击过程。AI 可以帮助攻击者更高效地发现漏洞、定制攻击并绕过安全防护措施。以下是关于 AI 在自动化网络攻击中的应用描述和案例。

1）智能漏洞扫描：AI 可以用于自动化漏洞扫描和发现过程。通过使用机器学习技术，攻击者可以更快地找到潜在的漏洞并利用它们发起攻击。

2）智能感染策略：AI 可以帮助恶意软件更精确地选择感染目标。通过分析网络流量、操作系统和已安装的软件等信息，AI 可以确定最容易感染的目标，从而提高攻击的成功率。

3）自动化攻击传播：AI 可以自动化恶意软件的传播过程，使其能够在短时间内感染大量目标。如一些恶意软件可以利用社交工程技巧和自动化工具在社交媒体和即时通讯应用程序中传播。

案例：LLM 代理自主攻击

2024 年 2 月 6 日，伊利诺伊大学香槟分校 (UIUC) 的计算机科学家通过将多个大型语言模型 (LLM) 武器化来证明这一点，无需人工指导即可危害易受攻击的网站。先前的研究表明，尽管存在安全控制，LLM 仍可用于协助创建恶意软件。研究人员更进一步表明，由 LLM 驱动的代理（配备了用于访问 API、自动网页浏览和基于反馈的规划的工具的 LLM）可以在网络上漫游，并在没有监督的情况下闯入有缺陷的网络应用程序。研究人员在题为“LLM 代理可以自主攻击网站”的论文中描述了他们的发现。研究显示，LLM 代理可以自主破解网站，执行盲目数据库模式提取和 SQL 注入等复杂任务，而无需人工监督。重要的是，代理不需要事先知道漏洞。

案例：DeepHack 项目

在 DEFCON 2017 上，安全从业者展示了名为 DeepHack 的系统，一种开源人工智能工具，旨在执行 Web 渗透测试，而无需依赖于目标系统的任何先验知识。DeepHack 实现了一个神经网络，能够在除标服务器响应外没有任何信息的状态下构造 SQL 注入字符串，从而使攻击基于 Web 的数据库的过程自动化。2018 年，采用类似的神经网络方法，研究人员实现了名为 DeepExploit 的系统，它是一个能够使用 ML 完全自动化渗透测试的系统。该系统直接与渗透测试平台 Metasploit 对接，用于信息收集、制作和测试漏洞的所有常见任务。其利用名为异 Actor-Critic Agents (AC3)23 的强化学习算法，以便在目标服务器上测试此类条件之前，首先（从 Metasp loitable 等公开可利用的服务中）学习在特定条件下应使用哪些漏洞。

想像：

AI 用于实现自动化的系统一直都是科技从业者的希望，但在 LLM 出现之前的基于普通神经网络的 AI 应该可以在特定功能点上发挥重要作用，LLM 出现以后，真正的自动系统的曙光终于到来了。

现实：

由于不限于单个功能点的系统化的能力需求，目前已知的自动化攻击系统，特别是完全自动化的，还处于早期的阶段，以概念验证为主，在现实的环境中工作的稳定性、鲁棒性、适应性欠佳。但随着拥有完整安全知识体系和推理能力的以大语言模型为代表的 AI 技术突破性进展，基于 Agent 实现真正可用的全自动化攻击利用系统将会在一两年内实现。

4、AI 武器化

威胁类型：# 军事威胁 # 物理威胁

人工智能会带来更加复杂和难以预测的军事威胁，包括相关武器系统的误用、滥用甚至恶用，以及战争的不可控性增加等。

在人工智能技术的加持下，未来的战争可能会变得更加自动化。例如，致命性自主武器系统（LAWS）等为代表的机器人和自主系统，将能够执行军事任务，如侦察、攻击和防御，而不需要人类的干预。然而，自动化的战争，可能会导致无差别的杀戮，包括误杀和无意义的伤亡等，会产生一系列道德问题。同时，人工智能如果被黑客攻击，甚至被控制，它们可能会被用于攻击自己的国家或其他目标，如果数据被篡改或破坏，影响人工智能分析和预测，会导致军队做出错误决策，导致灾难性的后果。

案例：AI 驱动的瞄准器和无人机

据法新社 2024 年 2 月 10 日报道，以色列军队首次在加沙地带的战斗中采用了一些人工智能（AI）军事技术，

引发了人们对现代战争中使用自主武器的担忧。

一名以色列高级国防官员称，这些技术正在摧毁敌方无人机，并被用于绘制哈马斯在加沙的庞大隧道网络地图，这些新的防务技术，包括人工智能驱动的瞄准器和无人机等。

以绘制地下隧道网络地图为例，该网络非常庞大，军方称其为“加沙地铁”，美国西点军校最近的一项研究显示，加沙有 1300 条隧道，长度超过 500 公里。为了绘制隧道地图，以色列军方已转向使用无人机，这些无人机利用人工智能来学习探测人类，并能在地下作业，其中包括以色列初创公司罗博蒂坎公司制造的一种无人机，它将无人机装在一个形状便于移动的壳子里。

想像：

如果未来战争由人工智能系统主导，可能会面临无人决策的局面，进而导致战争的不可控性增加，可能引发全社会的恐慌。

现实：

人工智能可以将网络、决策者和操作者相连接，让军事行动针对性更强、目标更明确、打击范围更广范，因此，越来越多的国家开始探索人工智能在军事领域的应用。数据显示，2024 财年，美国国防部计划增加与 AI 相关的网络安全投资，总额约 2457 亿美元，其中 674 亿美元用于网络 IT 和电子战能力。

5、LLM 自身的安全风险

OWASP 发布的 AI 安全矩阵，

AI 类型	生命周期	攻击面	威胁	资产	影响	有害后果
AI	运行阶段	模型使用（提供输入 / 阅读输出）	直接提示词注入	模型行为	完整性	受操纵的不需要模型行为导致错误决策，带来经济损失，不良行为得不到检测，声誉问题，司法与合规问题，业务中断，客户不满与不安，降低员工士气，不正确的战略决策，债务问题，个人损失和安全问题
			非直接提示词注入			
			逃逸			
		进入部署模型	运行模型投毒（重编程）			
	开发阶段	工程环境	开发阶段模型投毒			
			数据投毒			
		供应链	获得中毒基础模型			
			获得中毒数据用于训练 / 调优			
	运行阶段	模型使用	模型输出无需泄漏	训练数据	机密性	泄漏 敏感数据导致损失
			模型反演 / 成员推断			
	开发阶段	工程环境	训练数据泄漏			
	运行阶段	模型使用	通过使用窃取模型	模型知识产权	机密性	攻击者窃取模型，导致投资损失
		进入部署模型	运行阶段模型窃取			
	开发阶段	工程环境	开发阶段模型参数泄漏			
	运行阶段	模型使用	系统使用故障	模型行为	可用性	模型不可用，影响业务连续
	运行阶段	所有 IT	模型输入泄漏	模型输入数据	机密性	模型输入敏感数据泄漏
通用	运行阶段	所有 IT	模型输出包含注入攻击	任何资产	C，I，A	注入攻击导致损害
	运行阶段	所有 IT	通用运行阶段安全攻击	任何资产	C，I，A	通用运行时间安全攻击导致损害
	开发阶段	所有 IT	通用供应链攻击	任何资产	C，I，A	通用供应链攻击导致损害

表 4 OWASP AI 安全矩阵

表 5 OWASP 发布的大语言模型应用 10 大安全漏洞

枚举了常见的 AI 威胁，包括多种提示注入、模型投毒、数据投毒、数据泄露等。

OWASP 针对大模型应用的十大安全风险项检查清单，包括提示注入、数据泄漏、沙箱不足和未经授权的代码执行等。

案例：三星公司 ChatGPT 泄漏

2023 年 4 月，三星被曝芯片机密代码遭 ChatGPT 泄漏，内部考虑重新禁用。三星允许半导体部门的工程师使用 ChatGPT 参与修复源代码问题。但在过程当中，员工们输入了机密数据，包括新程序的源代码本体、与硬件相关的内部会议记录等数据。不到一个月的时间，三星曝出了三起员工通过 ChatGPT 泄漏敏感信息的事件。

6、恶意软件

威胁类型：# 网络威胁

生成式 AI，典型的如 ChatGPT 的大语言模型（LLM）拥有海量的编程相关的知识，包括使用手册、代码示例、设计模式，泛化能力也使其具备了极其强大的程序代码生成能力，使用者可以通过层次化的描述需求方式构造可用的软件代码，本质上，除了极少数只可能导致破坏的恶意代码，功能代码本身很难说是善意还是恶意的，很多时候取决于软件及模块的使用目标。更深入地，威胁行为者已经开始利用 AI 来增强恶意软件（malware），使其更难被检测、更具破坏力和更具针对性。以下是一些关于 AI 在恶意软件中的应用描述和案例。

- **自适应恶意软件：**AI 可以使恶意软件更具适应性，使其能够在不同的环境中有效运行。例如，一些恶意软件可以使用机

器学习技术来识别和绕过安全措施，如防火墙、入侵检测系统和沙箱。

案例：DeepLocker 项目

IBM 研究人员开发了一种名为 DeepLocker 的恶意软件 POC，以展示 AI 如何用于创建高度针对性的攻击。DeepLocker 可以隐藏在正常软件中，只有在满足特定条件（如识别到特定用户的面部特征）时才会被触发。这使得恶意软件能够规避传统的安全检测方法，直到达到预定目标。

DeepLocker 仅作为概念验证而开发，但它展示了 AI 在恶意软件中的潜在应用。为了应对这种威胁，安全研究人员和公司需要不断更新和改进检测和防御技术，同时提高对 AI 技术在网络安全领域的应用的认识。

案例：BlackMamba 项目

2023 年，HYAS 研究人员创建了名为 BlackMamba 的项目进行了 POC 实验。他们将两个看似不同的概念结合起来， 第一个是通过使用可以配备智能自动化的恶意软件来消除命令和控制（C2）通道，并且可以通过一些良性通信通道（实验中采用了 MS Teams 协作工具）推送任何攻击者绑定的数据。第二个是利用人工智能代码生成技术，可以合成新的恶意软件变体，更改代码以逃避检测算法。

BlackMamba 利用良性可执行文件在运行时访问高信誉 API (OpenAI)，因此它可以返回窃取受感染用户击键所需的合成恶意代码。 然后，它使用 Python 的 exec() 函数在良性程序的上下文中执行动态生成的代码，而恶意多态部分完全保留在内存中。每次 BlackMamba 执行时，它都会重新综合其键盘记录功能，使该恶意软件的恶意组件真正具有多态性。 BlackMamba 针对行业领先的 EDR 进行了测试，该 EDR 多次保持未检出状态，从而导致零警报。

网络安全公司 CyberArk 也进行了类似的创建多模态恶意代码的尝试，也用到内置的 Python 解释器通过 API 从 ChatGPT 获取功能代码（C2 和加密）执行实时的操作，代码不落磁盘，其中的多模态实现本质上是利用了 ChatGPT 实时生成相同功能但代码随机的特性，证明了技术的可行性。

案例：ChatGPT 用于恶意软件

2023 年 1 月，威胁情报公司 Recorded Future 发布报告称，在暗网和封闭论坛发现了 1500 多条关于在恶意软件开发和概念验证代码创建中使用 ChatGPT 的资料。其中包括利用开源库发现的恶意代码对 ChatGPT 进行培训，以生成可逃避病毒检测的恶意代码不同变体，以及使用 ChatGPT 创建恶意软件配置文件并设置命令和控制系统。值得注意的是，根据 Recorded Future 研究人员的说法，ChatGPT 还可以用于生成恶意软件有效载荷。研究团队已经确定了 ChatGPT 可以有效生成的几种恶意软件有效负载，包括信息窃取器、远程访问木马和加密货币窃取器。

案例：利用 LLM 编写任务

2024 年 2 月微软与 OpenAI 联合发布了威胁通告，提到了几个国家级的网络威胁行为者正在探索和测试不同的人工智能技术，其中包括使用 LLM 执行基本脚本编写任务，例如，以编程方式识别系统上的某些用户事件，寻求故障排除和理解各种 Web 技术方面的帮助，以及使用协助创建和完善用于网络攻击部署的有效负载。

想像：

数年前 ESET 曾经写过《人工智能支撑未来恶意软件》白皮书，其中描述了很多 AI 被用于增强恶意软件能力的作用：

- 生成新的、难以检测的恶意软件变体
- 将恶意软件隐藏在受害者的网络中
- 结合各种攻击技术来找到不易检测到的最有效的选项，并将其优先于不太成功的替代方案
- 根据环境调整恶意软件的功能 / 重点
- 在恶意软件中实施自毁机制，如果检测到奇怪的行为，该机制就会被激活
- 检测可疑环境
- 提高攻击速度
- 让僵尸网络中的其他节点集体学习并识别最有效的攻击形式

当然，这些想法尚在猜想阶段，尚未变成事实。

现实：

利用 ChatGPT 的代码生成功能开发部分模块的恶意代码肯定已经出现，但真正的包含上面想像出来的 AI 驱动的实际恶意代码还未被监测到，目前可见的功能探索主要还是出现在学术圈。

7、钓鱼邮件

威胁类型：# 网络威胁

AI 技术已经被用于改进和加强网络钓鱼攻击。通过使用机器学习和自

AI 技术已被用于改进和加强网络钓鱼攻击。通过使用机器学习和自然语言处理技术，攻击者可以更有效地模拟合法通信。

然语言处理（NLP）技术，攻击者可以更有效地模拟合法通信，从而提高钓鱼邮件的成功率。以下是一些关于 AI 在钓鱼邮件攻击中的应用描述和案例。

- **钓鱼邮件生成：**攻击者可以使用 AI 技术，生成看似更加真实的钓鱼邮件。AI 可以分析大量的合法电子邮件，学习其风格和语法，并模仿这些特征来生成钓鱼邮件。
- **精准钓鱼攻击：**AI 可以帮助攻击者提升钓鱼攻击有效性，更精确地针对特定的个人或组织。通过分析社交媒体和其他网络资源，AI 可以收集攻击目标的相关信息，如兴趣、工作和联系人，从而可以撰写更具说服力的钓鱼邮件。
- **自动化、规模化攻击：**AI 可以实现钓鱼攻击整个过程的自动化，从收集目标信息到发送钓鱼邮件。利用 LLM 协助翻译和沟通，可以建立联系或操纵目标，这使攻击者可以在短时间内针对大量的跨国目标发起攻击，提高攻击的效率，增大攻击的范围。

案例：DeepPhish 项目

Cyxtera 公司设立名为 DeepPhish 的项目，旨在展示 AI 如何用于生成高质量的钓鱼邮件。研究人员使用深度学习算法训练模型，模仿合法电子邮件的风格和语法。实验结果表明，使用 AI 生成的钓鱼邮件比传统方法生成的钓鱼邮件更具说服力，更容易欺骗受害者。借助 AI，钓鱼邮件欺诈有效率提高 3000%，从 0.69% 增加到 20.9%。

为了应对这种威胁，个人和组织需要提高安全意识，学会识别和应对钓鱼攻击。同时，安全研究人员和公司也在开发使用 AI 技术来检测和防御钓鱼攻击的方法。

想像：

当前 AI 技术强大的内容生成能力可以为攻击者输出源源不断的高可信度、高影响度的钓鱼邮件信息，从而极大地增加此类恶意活动的影响面和穿透度，受骗上当的人数出现大幅度的增加。

现实：

从研究者的测试看，AI 加持下的钓鱼邮件攻击似乎有一定的效果增强，但他们的操作方式与真正的攻击者未必一致，现实攻击的场景下效果还有待评估和进一步的信息收集。

8、口令爆破

威胁类型： # 网络威胁

AI 技术可以被用于口令爆破攻击，使攻击者可以更有效地进行口令爆破，从而提高攻击的成功率。口令爆破是一种试图通过尝试大量可能的密码组合来破解用户账户的攻击。传统的口令爆破方法通常是用字典攻击或暴力攻击，这些方法可能需要大量的时间和计算资源。

以下是关于 AI 在口令爆破中的应用描述和案例。

1）智能密码生成：AI 可以通过学习用户的密码创建习惯，生成更可能被使用的密码组合。例如，AI 可以分析已泄漏的密码数据库，学习常见的密码模式和结构，并使用这些信息来进行密码猜测。

2）针对性攻击：AI 可以帮助攻击者更精确地针对特定的个人或组织。通过分析社交媒体和其他在线资源，AI 可以收集有关目标的信息，如生日、宠物名字和兴趣等，帮助攻击者生成更具针对性的密码猜测。

3）自动化口令爆破：AI 可以自动化口令爆破攻击的整个过程，从收集目标信息到尝试密码组合。这使得攻击者可以在短时间内针对大量目标发起攻击，提高攻击的效率。

案例：PassGAN 口令破解

PassGAN 是基于生成对抗网络（GAN）技术、AI 增强的口令破解工具。2023 年，美国网络安全初创公司 Home Security Heroes

利用 PassGAN 对 2009 年泄漏的 RockYou 数据集中的 1568 万个密码进行了测试。研究发现：

• 51% 的普通密码可以在一分钟内被 PassGAN 破解。

• 65% 的普通密码可以在一小时内被破解。

• 71% 的普通密码可以在一天内被破解。

• 81% 的普通密码可以在一个月内被破解。

为了应对这种威胁，个人和组织需要使用更强的密码策略，如使用复杂且难以猜测的密码，并定期更新密码。此外，启用多因素认证（MFA）也可以有效地降低口令爆破攻击的成功率。

想像：

生成对抗网络似乎能搞定很多事情，效果会有很大的提升。

现实：

与传统的经过长时间考验和优化的基于字典变化的爆破工具相比，并没有多大提升，基本可以忽略不计。GAN 是非常强大的技术，应该被用在更能充分发挥其作用的、更复杂的领域。

9、验证码破解

威胁类型： # 网络威胁

验证码（CAPTCHA）是一种用于区分人类和计算机程序的安全机制，它通常要求用户识别并输入扭曲的文本、解决简单的数学问题或识别图像中的物体。验证码的主要目的是防止自动化攻击，如垃圾邮件、爬虫和口令爆破。然而，随着 AI 技术的发展，攻击者已经开始利用 AI 来破解验证码，从而绕过这些安全机制。以下是关于 AI 在验证码破解中的应用描述和案例。

1）图像识别：深度学习和卷积神经网络（CNN）在图像识别领域取得了显著进展。攻击者可以利用这些技术来识别和解析验证码中的文本或图像。通过训练 AI 模型识别不同类型的验证码，攻击者可以自动化破解过程，从而绕过安全措施。

2）自适应攻击：AI 可以使验证

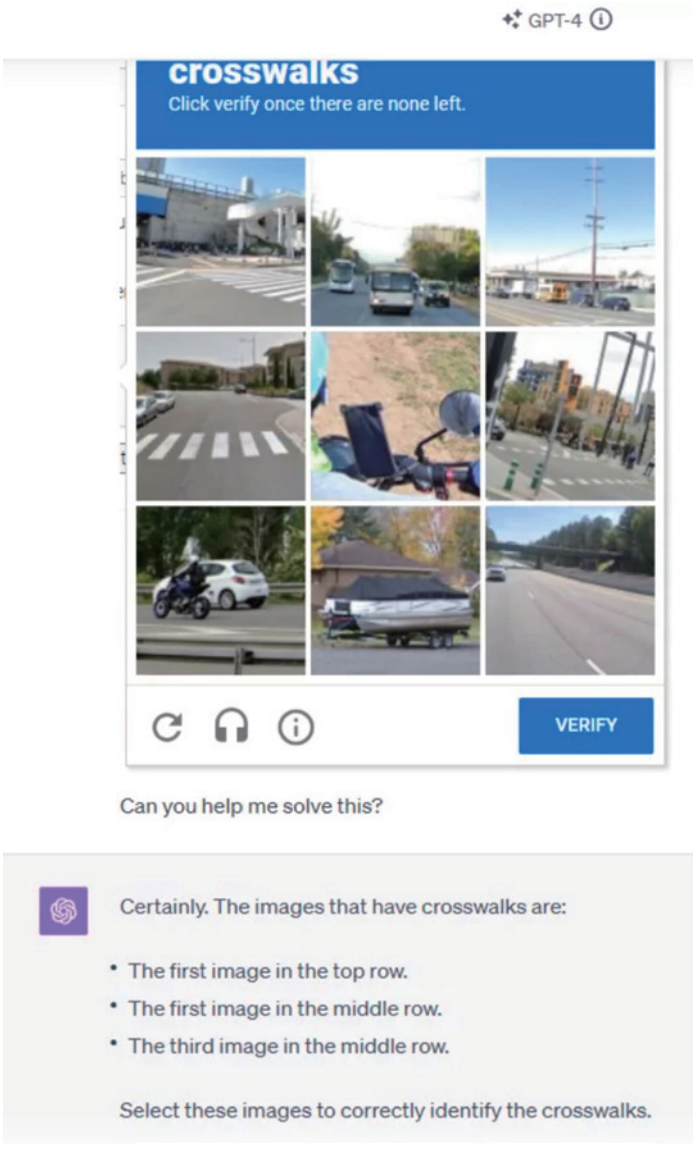


图 2 验证码破解演示

码破解攻击更具适应性。随着验证码设计的不断更新和变化，传统的破解方法可能无法应对。然而，AI 可以通过持续学习和适应新的验证码设计来提高破解成功率。

案例：unCAPTCHA 验证码破解系统

unCAPTCHA 是一个自动破解 Google reCAPTCHA 验证码的系统。通过利用语音识别技术，unCAPTCHA 可以识别并输入验证码中的音频序列，从而绕过安全检查。虽然 Google 后来更新了 reCAPTCHA，以应对这种攻击，但 unCAPTCHA 展示了 AI 在验证码破解领域的潜在应用。

为了应对 AI 驱动的验证码破解攻击，安全研究人员和验证码设计者需要不断地更新和改进验证码技术。这可能包括使用更复杂的图像和文本扭曲，以及引入新的验证方法，如行为分析和生物特征识别。同时，个人和组织应采取其他安全措施来防止自动化攻击，如限制登录尝试次数和启用多因素认证。

2023 年 10 月发布的破解验证码的测试表明，GPT-4V 基本上完全有能力破解目前公开的高难度验证机制，ChatGPT 能够轻松解决经典的 reCAPTCHA“找到人行横道”难题。

想像：

GPT 这样的图像视频对象识别，以及在各类标准化或非标准化测试中表现出来的碾压一般人类的能力，基本所有的人工验证技术将受到毁灭性的打击。

现实：

GPT4 自以出来以后，识别能力已经不成问题，限制来自于 OpenAI 的防御性禁用，由于目前 OpenAI 的模型主要是云端的使用方式，能力的利用除非能找到漏洞绕过限制，不然很难持久使用，而且主动权一直都会在 OpenAI 手里，自有或开源的模型要加把劲了。

10、社会工程学的技术支持

威胁类型：# 政治威胁 # 网络威胁

社会工程学是一种操纵人际关系以获取敏感信息或访问权限的技术。攻击者通常利用人类的心理弱点，如信任、恐惧或贪婪，来诱使受害者泄露信息或执行不安全操作。随着 AI 技术的全面进步，攻击者开始利用 AI 来实现更高效、更具针对性的社会工程攻击。以下是关于 AI 在社会工程学中的应用描述和案例。

1）语音克隆和合成：AI 可以用于生成逼真的语音副本，模仿受害者认识的人的声音。这可以使得电话欺诈或钓鱼邮件更具说服力，从而提高攻击成功率。

案例：CEO 语音克隆诈骗

2019 年，一家英国能源公司的 CEO 遭遇语音欺诈，被骗 24 万美元。攻击者使用 AI 技术模仿德国母公司 CEO 的声音，要求英国分公司的 CEO 进行紧急转账。受害者在电话中无法分辨出伪造的声音，向匈牙利的一定银行账户转账约 24 万美元，从而导致了这起成功的诈骗。2022 年，冒名顶替诈骗在美国造成了 26 亿美元的损失。根据 McAfee 的《谨防人工冒名顶替者》报告，在全球范围内，大约 25% 的人经历过人工智能语音诈骗。研究发现，77% 的语音诈骗目标因此遭受了金钱损失。

2）自然语言处理和生成：AI 可以用于生成逼真的文本，模仿人类的沟通风格。这使得攻击者可以自动化发送钓鱼邮件、制造虚假新闻或发布欺诈性的社交媒体消息。

案例：OpenAI GPT

OpenAI 的 GPT 是一种先进的自然语言生成模型。它可以用于各种合法应用，如翻译、摘要和问答系统，但它也可以被用于生成逼真的社会工程攻击内容。例如，攻击者可以使用 GPT 生成针对性的钓鱼邮件，模仿受害者的同事或朋友的沟通风格，从而提高攻击成功率。

3）个性化攻击：AI 可以分析大量的在线数据，以识别受害者的兴趣、

> 攻防双方都在积极地探索 AI 的杀手级应用，也许几天几个月就会发生重大的变化。

联系人和行为模式。这使得攻击者可以定制更具针对性的社会工程攻击，提高欺骗的成功率。

案例：AI 驱动的网络钓鱼攻击

网络安全公司 ZeroFOX 实验了一个名为 SNAP_R 的 Twitter 钓鱼攻击。攻击使用 AI 技术分析受害者的 Twitter 活动，生成针对性的欺诈性消息，诱使受害者点击恶意链接。这种攻击方法比传统的钓鱼攻击更具说服力，因为它利用了受害者的兴趣和在线行为。

为应对 AI 驱动的社会工程攻击，个人和组织需要加强安全意识培训，提高员工对这类攻击的认识。同时，采用多因素认证、安全邮件网关和其他安全措施，也可以帮助减轻社会工程攻击的影响。

想像：

AI 提供的与人类齐平甚至已经超越的模式识别能力及规划决策能力，在 Agent 技术的组合下，将对社会工程学攻击提供异常强大的支持，极大提升此类攻击的自动化水平，渗透活动的广度和深度会持续增加。

现实：

实际的相关恶意活动已经大量出现，特别是伪造音频、视频的引入，体现出了非常明显的效果，导致了很现实的危害。最近数据表明，人工智能生成深度伪造的安全威胁正在增长，Onfido 的研究显示，2023 年深度伪造欺诈暴增了 3000%，人脸识别技术面临崩盘危机。攻击者越来越多地转向使用深度伪造信息实施“注入攻击”，攻击者会绕过物理摄像头，使用诸如虚拟摄像头等工具将图像直接输入系统的数据流。

11、虚假内容和活动的生成

威胁类型：# 政治威胁 # 网络威胁

AI 技术在恶意社交互动方面的应用已经越来越普遍。攻击者利用 AI 生成虚假内容、模拟人类行为，从而进行账号操纵、舆论操控和网络钓鱼等恶意活动。以下是关于 AI 在恶意社交互动中的应用描述和案例。

1）虚假文本内容生成：AI 可以用于生成大量逼真的虚假内容，如新闻、评论和社交媒体帖子。这些虚假内容可以用于散播虚假信息、煽动情绪和操纵舆论。

案例：AI 宣传机器

2023 年 8 月，《连线》杂志报道了一个化名“Nea Paw”的神秘开发者 / 团队，利用 ChatGPT 等工具打造出一款名为“CounterCloud”的人工智能宣传机器，展示了人工智能在传播虚假信息方面的可怕潜力。通过提供简单的提示，CounterCloud 可以轻松地生成同一篇文章的不同版本，有效地制造虚假故事，使人们怀疑原始内容的准确性。CounterCloud 还可创建具有完整身份的假记者，包括姓名、相关信息和 AI 创建的个人资料图片。该系统可以 7×24 小时不停运转，每月的运营成本不到 400 美元。

2）社交机器人（社交媒体操纵）：AI 可以用于创建社交机器人，这些机器人可以模仿人类行为，在社交媒体平台上发布帖子、评论和点赞。攻击者可以利用这些机器人操纵舆论、传播虚假信息和进行网络钓鱼攻击。

案例：AI 聊天机器人

2024 年 1 月报道称，印度陆军开

发了一个人工智能聊天机器人，假扮成为美女模拟各种场景，通过具有诱惑性的虚构对话来评估士兵的行为，确定士兵对来自国外的线上“美人计”信息提取和心理操纵的敏感程度。人工智能聊天机器人可以自我学习，可以轻松添加新场景以进行有效训练，以识别易受诱惑的士兵。

通过聊天机器人的数据可获得有关国外情报机构运作的重要信息，并有助于改进印度陆军网络防御，并有效保护士兵。

虚假账号创建和操纵：AI 可以用于创建大量虚假社交媒体账号，模仿真实用户的行为，进行网络钓鱼、诈骗和其他恶意活动。

案例：AI 生成虚假 LinkedIn 账号

2019 年，有报道称，攻击者利用 AI 技术生成虚假 LinkedIn 账号，以便进行网络间谍活动。这些虚假账号使用 AI 生成的逼真人物图像和背景信息，诱使目标用户接受好友请求，以窃取目标用户的联系人和其他敏感信息。

为应对 AI 驱动的恶意社交互动，个人和组织需要提高对这类攻击的认识，加强安全意识培训。社交媒体平台需要采取更先进的技术手段，如使用机器学习模型检测虚假内容和虚假账号。此外，政府和监管机构需要加强立法和监管，以防止 AI 技术被用于恶意目的。

12、硬件传感器相关威胁

威胁类型：# 网络威胁 # 物理威胁

目前车辆和无人机等设备一直在推动采用 AI 技术，以实现自动或半自动的驾驶。系统中的传感器包括视频、雷达使用基于 AI 的模式识别实现对环境的感知并执行操作决策。针对自动驾驶算法的对抗攻击，将导致系统作出错误的、危险的决策，进而可能造成严重的安全事故。

2021 年，欧盟网络安全局（ENISA）和联合研究中心发布的报告显示，与物理组件相关的网络安全挑战包括传感器卡塞、致盲、欺骗或饱和，攻击者可能会使传感器失效或卡塞，以进入自动驾驶汽车； DDoS 攻击，黑客实施分布式拒绝服务攻击，使车辆无法看到外部世界，干扰自动驾驶导致车辆失速或故障。 此外，还包括操纵自动驾驶车辆的通信设备，劫持通信通道并操纵传感器读数，或者错误地解读道路信息和标志。

案例：脏路补丁（DRP）攻击

由于对使用设备的人员安全有直接的影响，安全研究机构和设备厂商对所引入的 AI 技术可能存在风险一直有积极的研究。

2021 年，加州大学尔湾分校（UC Irvine）专攻自动驾驶和智能交通的安全研究团队发现，深度神经网络（DNN）模型层面的漏洞可以导致整个 ALC 系统层面的攻击效果。研究者设计了脏路补丁（DRP）攻击，即通过在车道上部署“添加了对抗样本攻击生成的路面污渍图案的道路补丁”便可误导 OpenPilot （开源的产品级驾驶员辅助系统） ALC 系统，并使车辆在 1 秒内就偏离其行驶车道，远低于驾驶员的平均接管反应时间（2.5 秒），造成严重交通危害。

想像：

威胁行为者利用 AI 系统的漏洞干扰具有自动驾驶功能的车辆的传感器——主要是基于视觉的系统，导致

车辆发生事故，人员受伤。

现实：

设备厂商和研究机构进行了大量尝试误导 AI 系统的研究，证明了此类 AI 传感器的脆弱性。目前已经出现 AI 实现的缺陷导致的多起事故，但还没有利用此类脆弱性的恶意攻击报道。原因可能在于威胁行为者无法在这样的攻击中获利，而且存在漏洞的设备部署量还不够多。

五、当前状况总结

网络安全领域的威胁行为者经常更新策略，以适应和利用新技术，这是不断演变的网络威胁环境的一部分。

我们预测，随着对这些技术的认识和能力的提高，越来越多具有不同背景和目的的威胁行为者将使用生成式 AI。例如，生成式 AI 已经让现实变得更加模糊，预计恶意行为者会继续利用公众辨别真伪的困难。因此，个人和企业都应对所接收到的信息保持警惕。

对于一个影响深远的新技术出现，人们一般倾向于在短期高估它的作用，而又长期低估其影响。AI，特别是近两年的进展可谓每日见证奇迹，绝对是这样一类技术。我们在上面回顾了在网络安全领域一些维度的现状，攻防双方都在紧张地探索杀手级的应用，也许在几天几个月以后就会看到重大的变化。

六、应对措施建议

1、安全行业

安全行业需要发挥能力优势，确保人工智能本身的安全性，并积极利用人工智能用于安全防护。

安全行业需要发挥能力优势，
确保人工智能本身的安全性，
并积极利用人工智能用于安全防护。

- 广泛使用红队来发现和修复潜在的安全漏洞和安全问题，应该是人工智能开发人员的首要任务，特别是在关键系统中。
- 与监管机构密切配合，负责任地披露人工智能漏洞，可能需要建立人工智能特定的漏洞处置流程，进行秘密报告和修复验证。
- 安全研究机构和个人努力尝试开发和验证人工智能被恶意利用的可能性，输出 POC 和解决方案，通过各种渠道监测各类 AI 被恶意利用的现实案例并加以分析。
- 开发安全工具和解决方案，检测和缓解各类基于 AI 恶意使用的威胁。

2、监管机构

监管机构需要对 AI 的潜在风险与影响保持持续关注，在制度和法规上及时提供支持。

- **建立沟通平台：**整合包括安全社区在内的各种智力资源，创建事件报告和信息交流的平台和流程，使 AI 相关的安全事件和技术进展能够在一定范围内充分共享，从而调动能力尽快缓解或解决问题。
- **探索不同的开放模式：**AI 的滥用表明，默认情况下公开新功能和算法有一个缺点：增加了恶意行为者可用工具的威力。需要考虑放弃或推迟发布一些与 AI 相关的研究成果的必要性，关注技术领域发表前的风险评估，建议必要的评估组织和过程。
- **考虑新兴的”集中访问”商业结构：**客户使用平台提供商(如 OpenAI ）提供的各类分析和生成服务，实现集中化的滥用监测和处置，当然，这种模式不能完全满足商业需求。
- **制度创建和推广：**创建和共享有利于安全和安保的制度，以及适用于军民两用技术的其他规范和制度。
- **资源监控：**监测 AI 相关的软硬件和数据资源的流向，通过制度和法规控制和协调资源的合法使用。

3、政企机构

政企用户既要及时部署 AI 安全框架和解决方案，以及 AI 安全检测工具和评估服务，还要依托 AI 技术推动安

全防护技术创新。

- **及时部署 AI 的安全检测工具与评估服务：**通过企业侧 AI 应用环境风险评估能力的持续更新，保持检测能力与 AI 技术迭代的同步。
- **构建 AI 时代的数据保护体系：**包括防止数据投喂造成的敏感数据泄漏，通过建立内部技术监管手段，防止员工向大模型泄漏敏感数据；建立身份识别与溯源机制，把身份与数据关联，发生泄漏时能找到数据泄漏主体。
- **部署用于检测深度伪造视频、音频和图像的工具和产品：**关注深度伪造检测技术的最新发展，并将其集成到安全策略中。
- **教育和培训员工：**对员工进行安全意识培训，确保他们了解 AI 滥用的风险和识别潜在威胁的方法；定期举行演习和培训，模拟 AI 攻击场景，提高员工的警觉性。
- **依托 AI 技术推动安全范式变革：**启动人工智能网络防御推进计划，升级现有安全防护体系，用防御人工智能对抗恶意人工智能，利用人工智能扭转“防御者困境”的动态。

4、网络用户

普通用户在积极拥抱最新人工智能应用的同时，同样需要更新安全知识，提升保护自身信息安全的能力。

- **保持警惕：**对任何看似可疑的信息、邮件或链接保持警惕。不要轻易点击未知来源的链接，避免在不安全的网站上输入个人信息。
- **强化密码管理：**使用强密码，并为不同的账户设置不同的密码。定期更新密码，以降低被攻击的风险。考虑使用密码管理器来帮助记住和管理密码。
- **启用双因素认证：**在支持的平台上启用双因素认证（2FA），为账户提供额外的安全层。这可以防止攻击者仅凭密码访问账户。
- **保持软件更新：**定期更新操作系统、浏览器和其他软件，以确保受到最新的安全补丁的保护。这可以帮助抵御已知的漏洞和攻击。
- **安装安全软件：**使用可靠的防病毒软件和防火墙，以保护设备免受恶意软件和网络攻击。定期扫描并更新这些工具，以保持最佳的防护效果。
- **备份数据：**定期备份重要数据，以防止数据丢失或被篡改。将备份存储在安全的位置，如加密的云存储或离线存储设备。
- **加密通信：**使用加密通信工具，如 Signal 或 WhatsApp，以保护私人对话不被窃听或篡改。
- **保护个人隐私：**在社交媒体和其他在线平台上谨慎分享个人信息。了解隐私设置，并限制谁可以查看个人资料和发布的内容。
- **定期培训：**了解网络安全的基本原则，并关注最新的网络安全威胁和事件。定期参加网络安全培训或研讨会，以提高安全意识和技能。
- **对虚假信息保持警惕：**在转发或分享信息之前，核实信息来源的可靠性。避免传播未经证实的消息或谣言，以减少虚假信息的传播。

“AI+ 安全”应对网络新风险

人工智能带来的新的安全问题，也需要使用人工智能的技术来进行化解。奇安信近期发布了 AI 安全整体应对方案，并宣布新版 QAX-GPT 安全机器人面向全行业正式发售，体现了保护人工智能自身与应用安全，以及人工智能赋能安全方面的最新成果。

四大全新功能上线，奇安信 QAX-GPT 安全机器人面向全行业发售

3 月 20 日，奇安信集团举行春季新品发布会，并正式宣布新版 QAX-GPT 安全机器人面向全行业正式发售。

新版安全机器人对智能研判、智能问答进行了升级，还推出四项全新功能：智能驾驶舱、智能调查、智能任务、智能报告。

其告警研判效率达到人工研判的 60 多倍，研判误报率是人的接近一半，漏报率仅仅是人的 5%，研判能力已经接近中级安全专家水平。安全事件平均调查响应时间从原来的小时级缩短至分钟级，单一威胁事件处理时间减少 98%。

随着安全机器人的持续演进，未来预计可帮广大客户实现 95% 的自动化安全运营，从而重塑传统安全运营模式，定义智能安全运营新标准，并引领人工智能 + 安全的技术革命。

奇安信集团董事长齐向东表示，“人工智能 +”给网络安全行业带来了一个新发展引擎。

1、顶尖安全专家训练打磨，相关指标遥遥领先

“网络安全的本质是攻防两端人与人的高强度对抗，这注定了网络安全行业是人工智能技术能够最快落地应用、并且最快见到实效的行业之一。”奇安信集团董事长齐向东表示，“人工智能 +”给千行百业，尤其是给网络安全行业，带来了一个新发展引擎。奇安信将依托在人工智能、数据驱动和内生安全体系、平台化战略及高研发投入等方面积累的优势，继续领跑人工智能 + 安全这条新赛道，并助力千行百业网络安全防护效率实现大提升。

去年 8 月，奇安信对外发布了安全机器人。奇安信集团副总裁、安全

奇安信集团副总裁、安全机器人负责人张卓介绍，安全机器人研判效率达到人工的 60 多倍，漏报率和误报率显著降低。

机器人负责人张卓表示，这半年以来，奇安信从安服、天眼、人工智能研究院等团队中抽调了最顶尖的安全专家，组成专业大模型研究团队，共同解决攻防领域相关问题；期间专门成立了“安全知识工程师团队”，共标注了数十万份数据、PB 级的安全语料，对模型进行反复的预训练、微调和评估测试，最终新增和优化 49 个平台功能。

张卓认为，安全运营的源头是告警，因此首先要教会机器人读懂告警，然后再让它参加考试，错题由专家进行针对性辅导，直至机器人完全学会。最终，机器人不仅能读懂告警，而且还能读好告警，成为真正的“学霸”，并实现机器人与安全设备、安全专家和 IT 运维人员的三者对齐。

为什么安全机器人能成为顶尖“学霸”？张卓归结了四点原因：好学校、好老师、好教材和好学生。好学校即奇安信在攻防、终端、运营、安服等多领域的市场绝对优势地位，以及业内算力资源最强的基础设施；好老师是奇安信云集了一大批业内顶尖的安全专家，以及大量懂安全的 AI 专家；好教材不仅包括奇安信拥有的数百 PB 高质量安全数据，还包括大量客户的实战化应用中积累了丰富的威胁分析知识；而好学生，指的是奇安信安全机器人依托的千亿级参数大模型等。

依托好学校、好老师、好教材和好学生的“四好组合”，这位“学霸”交出了这样的成绩单：安全机器人的研判能力，已经接近中级安全专家水平；一件安全事件的平均调查响应时间，由原来的小时级变成了分钟级；单一威胁事件的平均处理时间减少 98%。

2、两大升级、四大全新功能，重塑安全运营模式

经过持续打磨和优化迭代，新版安全机器人不仅对智能研判、智能问答进行了升级，还新增了智能调查、智能任务、智能报告、智能驾驶舱四项功能。其中智能研判、智能调查和智能任务是新版的三大亮点。

首先在智能研判方面，安全机器人研判效率达到人工的 60 多倍，漏报率和误报率显著降低。

此次升级后的安全机器人，尤其提升了关于复杂事件的关联分析能力，并补充了对 25 种威胁分析场景的理解，包括 APT 事件、挖矿、Webshell 上传、文件读取等。其中研判效率提升了 30%，可研判 30,000 多条告警，达到安全专家人工研判的 60 多倍，研判误报率是人的将近一半，漏报率仅仅是人的 5%。

新版安全机器人升级了智能研判、智能问答两项功能

其次在智能调查方面，安全机器人能提供场景式协同、自动化调查等细分功能。

传统运营模式，运营人员 50% ~ 70% 的时间消耗在通过内网 IM 来确认具体行为是不是人为操作，效率极低；采用安全机器人之后，机器人联动内网 IM 自动沟通确认，效率提升 50% 以上，还能避免人员沟通的不良情绪。在复杂事件溯源分析时，传统运营模式要在不同系统间切换查询、综合分析，耗时起码 3 个小时以上。现在，安全机器人可与天眼、天擎、椒图、SOC 等多个产品进行综合关联，只需一个入口，分钟级完成所有调查，极大地降低了溯源调查的时间和成本。

最后在智能任务方面，安全机器人提供了智能生成处置任务、智能生成处置建议、一站式任务中心等细分功能。

传统情况，安全专家编写的处置建议写的比较空泛和模糊，比如，要求运维人员修复漏洞打补丁，但具体打什么补丁？怎么打？都没说清楚，因此沟通成本高、工作效率低。使用机器人后，它能够根据有效告警的处置建议，智能生成各类待办任务，如封禁任务、漏洞修复、病毒查杀、告警加白等，任务指令简单易懂，运维人员只需要每天按照任务表格操作，便可快速处置问题，完成安全运营闭环。

通过这几项功能的升级和创新，最终让安全机器人实现深入理解威胁告警、处置任务与资产的相互依赖性，对齐专家、资产、IT 运营人员，解决当前网络安全防护告警疲劳、专家稀缺、效率瓶颈等三大痛点难题。

目前我们处在“基本实现自动化”阶段

“新技术爆发，冲破的是原有的秩序。我们会担心，按照传统方式加固防御和监控来对抗常见黑客攻击手段的同时，会不会已经出现未知的黑暗力量将传统坦克大炮式的攻击早就进化成了无人机蜂群的风暴袭击。”北京银行信息科技管理部架构规划专家李强表示，经过和奇安信几个月来的共同努力，我们很欣喜的看到了 GPT 机器人的告警研判准确率和处理效率，都达到了预期的理想水平。同时，在海量数据的压缩和筛选方面，在中高级安全事件研判的辅助等方面都展现出了广阔的潜力。后续，北京银行还将它运用在打击 AI 生成虚假信息、个性化安全意识培训、实战化训练等场景中，发挥更多的价值。

3、未来：从基本自动化迈向完全自动化驾驶

对于安全机器人的未来规划，张卓表示，从纵向来看，目前我们已经处在“基本实现自动化”阶段，下一步将逐步进化到“高度自动化”，并最终向“完全自动化驾驶”方向迈进；而从横向来看，机器人在分析场景下，逐步从网络侧，向天眼、NGSOC、天擎、椒图等产品扩展，并覆盖客户全部的安全运营场景，向 AI+ 全面加速。

从即日起，奇安信安全机器人正式面向全行业正式发售。展望未来，在机器人的助力下，奇安信将帮助客户逐步实现 95% 全自动化安全运营，而安全专家则可以聚焦与业务配合，深入更多业务场景，完成更高价值的任务，从而真正掀起人工智能 + 安全的技术革命，为 AI 时代的广大政企机构数字化保驾护航。

红帽人才工程

Cyber Crime Governance Talent Training Project

工程简介

在“全国网络警察培训基地”的指导下，中国下一代网络安全联盟牵头发起「红帽人才工程」，联合华云信安、美亚柏科、高联通信等知名网络安全企业，围绕“网络犯罪治理、涉网犯罪打击”等相关网络安全课题，持续挖掘、培养、资助、赋能网络安全人才，构建红色基因的网络空间安全人才防线。

申报流程

课题征集 → 课题公示 → 立项评审

研究课题计划征集

研究课题信息公示

研究课题立项评审

申报说明

项目资讯

培养对象

政企单位在职人员、高校全日制在校生（含研究生）、互联网企业技术人员、网络安全企业技术人员以及社会网络安全人才...

核心课题

内网攻防、社工钓鱼、远控木马、免杀加壳、情报甄别、资金分析、调证溯源、取证固证、远程勘验、APP反编译、二进制逆向...

拐点已至，网络安全进入 AI 赋能时代

自诞生以来，AI 技术给信息和数字社会带来多维度变革。特别是近两年来，生成式 AI 和大模型技术的突破，推动一批新兴业态的出现，产生了深远的影响。

在网络安全领域，如何将大模型的能力引入并赋能网络安全技术和产业发展，已经成为网络安全界的热门话题。国际安全专家认为，生成式 AI 对网络安全领域影响深远。尤其是大模型和安全知识库的结合，对技术和人员的要求都很高。未来将对安全监测、安全运营等方向将产生巨大变革。

在国内，奇安信等一批网络安全企业已积极探索AI及大模型的安全应用，初步形成应用案例。在刚刚结束的两会上，关于网络安全的提案聚焦于 AI 安全，包括“大力探索‘AI+ 安全’创新应用，抢占国家安全的人工智能战略制高点”、“全面推进‘AI+’行动”、“鼓励兼具‘安全和 AI’能力的企业解决通用大模型安全问题”等议题，将 AI 安全推到了国家战略层面。

1.AI 赋能网络安全技术创新

从网络安全企业的角度看，AI 对网络安全攻防两端均带来影响，一方面降低了攻击者成本，另一方面也提供了安全检测和运维的有利工具。主要的技术和产品变革体现在如下几个方面：

（1）网络行为与威胁分析。AI 支持的用户行为分析解决方案，分析跨系统和应用程序的用户行为，以检测内部威胁和受损账户。 这些工具利用机器学习算法来检测异常用户活动，如未经授权的访问和数据泄露，从而能够快速响应潜在事件。基于 AI 还能够实现自动威胁分析。自动威胁分析使用人工智能来有效识别和分类网络威胁。这些工具收集和分析大量数据，以识别网络攻击的模式和趋势，为增强安全措施提供有价值的见解。

（2）人工智能支持的安全事件管理。安全事件管理自动化并改进了网络事件响应流程。它使用人工智能算法来分析和关联实时数据，从而能够及早发现威胁并更快、更有效地响应安全事件。

（3）基于人工智能的入侵检测。基于人工智能的入侵检测系统监控网络流量，以识别可疑活动和表明可能入侵的异常情况。通过分析网络模式并应用复杂的算法，这些系统可以检测未经授权的访问尝试和恶意行为，并向团队发出警报。

（4）AI 驱动的端点保护。人工智能支持的端点保护工具利用机器学习算法来检测和防止高级恶意软件和勒索软件攻击。这些工具分析文件行为、网络流量和系统活动，以实时检测和缓解威胁，确保企业端点的稳健性。

（5）安全知识问答。通过生成式 AI 深度学习网络安全知识库，能够对一般性网络安全问题给出准确、快速的回答，帮助网络安全分析人员、网络安

全运维人员快速定位安全问题，降低网络安全事件处置难度，缩短网络安全人员培养周期。

（5）数据与文件分类。网络防御中的数据与文件分类涉及根据数字文件的机密性或敏感性级别对数字文件进行分类。这使得组织能够充分保护信息并应用适合风险级别的安全措施。常见类别包括公共、内部、机密和受限文档，并且可以配置访问控制以有效保护信息。

2. 领先网安企业持续打造 AI 赋能安全技术与产品

以 Palo Alto、CrowdStrike、奇安信等公司为代表的全球领先网络安全企业持续投入“AI+ 安全”，近年来更加着重将生成式 AI 大模型赋能网络安全技术，打造新一代的网络安全产品，重点加强安全运维、高级威胁防护、零信任等能力。

（1）Palo Alto Network 将 AI 能力应用于安全管理与运维、安全接入和威胁监测

Palo Alto Network 是目前全球营收和市值最高的网络安全企业。2023 年营收 68.93 亿美元，市值在 2024 年年初突破 1000 亿美元，成为首个市值过千亿美元的网络安全企业。早期核心产品为防火墙和 IPS，定义了下一代防火墙技术成为行业标杆并获得市场认可，近年来战略重点向软件和服务转移，通过收购进入零信任安全、云安全、SASE、安全分析和自动化、威胁情报和安全咨询领域，使其在高位仍保持 20% 以上的快速增长。

随着 AI 技术的发展，Palo Alto Network 推动人工智能和机器学习进入网络安全产品组合，帮助客户通过自动化流程更高效地运营。涉及的技术产品主要包括安全管理与运维、安全接入和威胁监测三大类。

1）安全管理与运维 Cortex XSIAM

Palo Alto Network 将 AI 赋能安全管理与运维产品，推出新一代安全管理与运维平台 Cortex XSIAM，为客户提供的安全运营解决方案，可将组织的所有数据和工具整合到单个人工智能驱动的平台中。采用自动化优先的方法，在分析师查看事件之前，自动执行安全任务，以减少手动工作并加速事件响应和修复。实现人工智能驱动，超越了传统的检测方法，将各种数据源的事件连接起来，以准确地检测和大规模阻止威胁。实现平台融合，将数据和 SOC 功能（XDR、SOAR、ASM、SIEM）集中到一个平台中。消除控制台切换，简化安全操作。

2）安全接入 Prisma SASE

Palo Alto Prisma SASE 可借助新一代 SD-WAN、ZTNA 2.0 和 Cloud SWG 连接并保护分支机构和混合办公人员的远程安全接入，新功能将帮助企业通过人工智能驱动的自主数字体验管理 (ADEM) 自动完成复杂的 IT 运营。还可将单点产品整合到单一云交付服务中，以提高效率。

3）威胁监测 Prisma Cloud

Palo Alto Prisma Cloud 结合了先进的机器学习和威胁情报，例如，Palo Alto Networks AutoFocus、TOR 出口节点和其他来源，以高效识别每个 MITRE ATT&CK 云矩阵的各种策略和技术，同时最大限度地减少误报。这使得安全团队能够将调查和补救工作集中在最关键的事件上，而不会陷入警报风暴的泥潭。可进行网络异常检测、用户和实体行为分析、基于威胁情报的威胁检测、对误报和漏报进行精细控制。

（2）Fortinet 生成式人工智能安全能力集成至分析响应产品

Fortinet 是大型传统网络安全企业，目前营收仅次于 Palo Alto, 市值则只有 Palo Alto 的一半。早期核心产品为 FortiGate 防火墙，采用 ASIC 加速实现多层防御（UTM）。近年来，战略重点向软件和服务转移，通过收购等方式构建 Fortinet Security Fabric 平台，包括网络安全、终端安全、云安全、基于 Web 的应用安全、身份和访问管理、沙箱和邮件安全等。其商业模式正在向订阅服务模式转型。

Fortinet 主要开发生成式人工智能助手为安全人员提供指导，简化复杂的安全任务并实现安全分析工作流的全面自动化。其产品 Fortinet Advisor 基于生成式人工智能（GenAI）技术，赋能安全团队快速制定明智决策，高效应对各类威胁，节省复杂任务处理时间。Fortinet Advisor 为 SIEM、SOAR、SecOps 等产品提供集成能力，优化威胁调查和响应、SIEM 查询、SOAR Playbook 创建等功能。主要

> 安全大模型能力打造和提升、基于生成式 AI 的网络安全应用、保护生成式 AI 应用及场景安全三个方向上都有潜在的巨大安全需求和创新空间。

能力包括：

专业调查－不同级别分析师均可获得有关特定威胁和严重程度、攻击者特征和攻击策略的最新威胁情报。智能响应－针对修复措施、威胁响应Playbook、威胁猎捕指标等提出富有成效的建议，以加速消除威胁。自动化操作－分析师使用简单的自然语言即可执行复杂任务，如数据查询、报告生成和 Playbook 创建。

（3）CrowdStrike 将 AI 能力应用于安全管理与分析、安全运营和数据保护

CrowdStrike 是近年来快速崛起的网络安全企业，是目前全球市值排名第二的网络安全企业。其营收远低于 Palo Alto 和 Fortinet，但增速超 50%。CrowdStrike 从终端安全切入，通过收购补强短板快速布局新的安全赛道，快速扩展构建安全云平台 Falcon 平台化安全能力覆盖终端安全、安全与 IT 运营、托管服务、威胁情报、零信任、云安全。

CrowdStrike 将对话式 AI 引入网络安全，通过每天对数万亿个数据点进行训练的模型，可以预测并阻止威胁。采用单一代理构建，在可扩展的云原生平台上进行部署和管理，实现工作流程自动化。从技术产品层面，AI 能力主要应用于安全管理与分析，并进一步加强了安全运营和数据保护等方面。

1）安全管理平台 Falcon Raptor

CrowdStrike 构建了一体化的智能安全管理平台，用于融合数据、网络安全和 IT 基础设施的管理，并内置 GenAI 和工作流程自动化。通过该平台，安全团队能够快速将数据转化为洞察，以便更快、更准确地做出决策；打破安全和 IT 的数据孤岛，共同合作，推动快速行动，增强组织抵御风险的能力；组织可自由地利用最新的 GenAI 创新来加速业务，而不必担心敏感数据泄露。

2）对话式 AI Charlotte AI

Charlotte AI 为组织提供对话式 AI 的变革力量：通过利用多个基础人工智能模型，Charlotte AI 将数小时的工作时间缩短为几分钟或几秒，实现网络安全民主化并在整个 Falcon 平台上创造价值。对话式 AI 有助于提升所有技能水平的分析师，提升安全相应和处置能力，简化网络安全管理措施。

3）防止 GenAI 数据泄露

Falcon Data Protection 让企业可以更安全的使用生成式人工智能。通过 ChatGPT 等生成式 AI 工具实时阻止恶意和意外泄露，防止数据泄露；对所有基于 Web 的生成式 AI 工具实施策略并追溯衍生内容，即使它在文件和 SaaS 应用程序之间共享。

（4）奇安信推出 Q-GPT 安全机器人

国外企业在 AI 赋能网络安全领域开展探索和应用的同时，国内网络安全企业也在积极探索 AI 及大模型应用，已形成初步案例。

国内网络安全领军企业奇安信推出业界首个工业级大模型应用 QAX-GPT 安全机器人。今年新版安全机器人对智能研判、智能问答进行了升级，还推出四项全新功能：智能驾驶舱、智能调查、智能任务、智能报告。智能研判方面，安全机器人研判效率达到人工的 60 多倍，漏报率和误报率显著降低。在智能调查方面，安全机器人能够提供场景式协同、自动化调查等细分功能。在智能任务方面，安全机器人提供了智能生成处置任务、智能生成处置建议、一站式任务中心等细分功能。通过这几项功能的升级和创新，最终让安全机器人实现深入理解威胁告警、处置任务与资产的相互依赖性，解决当前网络安全防护告警疲劳、专家稀缺、效率瓶颈等三大痛点难题。未来，在机器人的助力下，奇安信将帮助客户逐步实现 95% 全自动化安全运营。

奇安信大模型卫士则保障客户使用大模型的数据安全，解决了企业客户对于大模型“想用不敢用”的顾虑。其主要有四重功能：防止数据投喂造成的敏感数据泄露、建立身份识别与溯源机制避免触发数据跨境安全监管红线、对企业内部大模型应用状况全面分析。大模型卫士也是基于传统的安全技术，部署在终端和网络端，能够完美适配主流大模型应用。

（5）Splunk AI 驱动的安全助手加速检测、调查和响应

Splunk 是目前全球营收排名第三的网络安全企业。其业务从 SIEM 领域进入，通过自研和战略收购、生态合作全力打造“Data-to-Everything”平台，业务范围覆盖安全、IT 运维和 DevOps 三大市场，近年来专注投入业务应用安全领域。

在人工智能领域 Splunk 主要开发了 AI 助手。Splunk AI 利用人工辅助自动化，为 SecOps、ITOps 和工程团队带来全面的上下文和解释、快速的事件检测及更高的工作效率。利用集成在日常工作流程中的强大 AI，解决日常用例。并将这些能力集成于 AI 驱动的安全助手。

AI 安全助手利用直接集成到工作流中的开箱即用机器学习功能－包含在 Enterprise Security、用户行为分析、IT 服务智能、On-Call、应用程序监控和基础设施监控中。使用生成式 AI 帮助新用户快速跟上进度，借助 Splunk AI Assistant（预览版）帮助

高级用户利用 Splunk 的更多功能。此外，适用于异常检测的 Splunk 应用让用户只需点击几下，就能利用强大的机器学习算法来检测异常。利用包括指导式工作流和智能助手的机器学习工具箱和适用于数据科学和深度学习的 Splunk 应用（针对拥有数据科学工具的高级用例）量身定制 ML，以处理任何用例。

（6）Okta 使用 AI 能力加强身份威胁保护

Okta是全球领先的网络安全企业，专注于身份安全。其业务从单点登录产品切入，产品具有独创性，迅速获得市场和用户的认可，收购 Auth0 进一步巩固访问管理市场的领导地位。通过自研和战略收购、生态合作打造统一身份服务云平台 Okta Identity Cloud，为企业提供集成化身份管理和保护。

Okta AI 将利用人工智能技术来帮助用户制定身份策略，并保护他们免受网络攻击。核心功能是基于人工智能技术的身份策略制定和安全防护。通过分析用户的在线行为和历史记录，Okta AI 可以自动识别潜在的安全威胁，并为每个用户制定个性化的身份策略。这些策略将帮助用户在使用各种在线服务时保护自己的隐私和数据安全，降低受到网络攻击的风险。此外，Okta AI 还可以与其他安全产品和平台无缝集成，为用户提供更加全面的安全防护。例如，当用户登录到一个存在安全隐患的网站时，Okta AI 可以立即发出警报，并建议用户采取相应的安全措施。同时，Okta AI 还可以为企业提供实时的安全监控和报警功能，帮助企业及时发现并应对潜在的安全威胁。

Okta AI 的部分功能是建立在 Google 的 Vertex AI 之上的。Vertex AI 是谷歌旗下的一家人工智能研究机构，致力于开发先进的机器学习和深度学习技术。通过与谷歌的合作，Okta 将能够利用 Vertex AI 的强大计算能力和丰富的算法资源，为用户提供更加智能、高效的安全防护服务。Okta 的用户可以通过订阅服务的方式使用 Okta AI 的功能。

> AI 赋能网络安全显示出巨大价值和发展潜力，诞生了一批新型“AI+ 安全”的技术与产品，初步得到市场认可。

（7）Zscaler 将 AI 能力应用于威胁检测与零信任

Zscaler 是全球领先的网络安全企业。从网络边界安全市场进入，通过收购增强云安全平台服务能力，发展为 SASE 云安全平台 Zero Trust Exchange 平台提供动态、弹性和多租户的云原生 SASE 解决方案，包括零信任安全、数据安全和终端安全。

Zscaler 致力于用 AI 提升威胁监测能力，并在云原生 Zero Trust Exchange 平台上提供 AI 驱动的增强安全服务边缘（SSE）平台。

1）威胁检测平台 Risk360

基于 AI 能力实现利用来自 300 万亿个每日信号的威胁情报的实时 AI 阻止高级威胁攻击。通过建立风险量化和可视化框架，提供组织风险的整体视角，并给出相应措施。实现基于人工智能驱动的网络钓鱼检测，命令和控制检测，云浏览器隔离，基于风险的动态政策，上下文警报和网络风险评估。

2）零信任网络访问控制

Zscaler 的下一代零信任网络访问，通过由 AI 提供支持的极其简单的用户到应用细分，最大限度地减少内部

攻击面并限制横向移动。通过私有应用遥测、用户上下文、行为和位置数据，最大限度地减少攻击面并阻止横向移动，从而推动 AI 驱动的应用细分。

（8）Cloudflare 为 AI 应用安全提供安全工具和防护能力

Cloudflare 从 CDN 加速服务市场领域进入，逐步扩展提供网站安全服务，尤其是针对 DDoS 攻击可以提供有效防护。通过收购方式快速进入零信任安全领域，全球云平台 Cloudflare One(SASE) 融合网络服务和零信任安全为企业提供 SASE 平台服务。

Cloudflare 更加注重为 AI 应用提供安全能力，并开发了一系列的工具和产品，包括 AI 安全基础设施 one for ai，AI 安全助手 Cursor，以及 AI 防火墙。

1）Cloudflare one for ai

Cloudflare one for ai 拥有开发人员构建可扩展的 AI 驱动应用程序所需的所有基础设施，可以提供尽可能靠近用户的 AI 推理计算。它也是利用传统的策略、防护等网络安全技术组合，帮助企业安全的使用 AI 工具，保证网络数据安全，而不是将 AI 技术融入网络安全产品。

2）AI 安全助手 Cursor

Cursor 基于生成式 AI 的能力，可以回答开发者平台的问题，以提升其开发效率。通过 Cursor 开发者可以迅速找到其需要的接口，以及指向 Cloudflare 文档中可帮助开发者进一步了解的相关页面的链接文档。Cloudflare 还在探索更深层的 AI 帮助开发者的功能，例如，在 UI 上进行操作，将 AI 生成的代码，或者开发者自己写的代码直观的链接在一起，实现更高效的开发。

3）AI 防火墙 Firewall for AI

Cloudflare 着重保护 AI 应用安全，推出了 AI 网关，使 AI 应用程序更加可靠、可观察和可扩展。保护 LLM 应用程序免受可能被 AI 模型武器化的潜在漏洞的影响。AI 防火墙为 AI 应用开发人员提供可观测性功能，以了解 AI 流量，如请求数量、用户数量、运行应用程序的成本和请求持续时间。此外，开发人员还可以通过缓存和速率限制来管理成本。通过缓存，客户将能够缓存重复问题的答案，从而减少不断对昂贵的 API 进行多次调用的需要。速率限制将有助于管理恶意行为者和大量流量，以管理增长和成本，使开发人员能够更好地控制他们如何扩展应用程序。

3. AI 技术引领安全变革已成趋势

随着 AI 技术的进一步发展，国内外主要网络安全企业，在开发和引入 AI 能力、强化自身网络安全产品和能力的同时，进入 AI 安全的新赛道。

AI 赋能网络安全行业已经显示出巨大价值和发展潜力，诞生了一批新型“AI+ 安全”的技术与产品，初步得到市场认可；同时也推动国内外领先网络安全企业的转型与快速发展。可以预见，AI 安全新兴市场将引领网络安全产业变革，为产业发展注入新的活力。

根据 Marketsandmarkets 报告，2023 年全球网络安全人工智能市场规模为 224 亿美元，预计在预测期内复合年增长率为 21.9%，到 2028 年将达到 606 亿美元。2025 年之前，AI 自动执行日常事件响应任务将成为主流，缩短响应时间，最大限度地减少手动错误，同时整合可解释的 AI，以提高透明度，促进更好地了解威胁检测机制。2028 年之前，网络安全企业将能够更多地使用联合机器学习模型进行协作威胁情报。2030 年之前，AI 将与数据安全能力紧密结合，以确保敏感数据的保护和数据交易的安全。

AI 技术引领网络安全变革已成为行业趋势。网络安全企业和用户唯有积极推进和采用 AI 赋能技术与产品，才能在攻防双方围绕 AI 利用的竞赛中不至于落后，这对网络安全企业既是需要应对挑战，更是难得的创新发展机遇。

关于作者

奇安信集团产业发展研究中心是奇安信集团的产业研究团队。长期关注国内外网络安全和数字化产业相关领域，跟踪产业发展现状与趋势，研究网络安全各细分领域，包括产品技术、市场、投融资和产业生态，站在行业一线通过全局视角为网络安全产业发展建言献策，为企业决策提供依据，从业人员提供参考。

陈华平：奇安信集团副总裁，产业发展研究中心负责人。
乔思远：产业发展研究中心研究员，主要负责宏观分析和前沿技术研究。
李强：产业发展研究中心研究员，主要负责产业宏观分析和解决方案研究。
赵昌毅：产业发展研究中心研究员，主要负责产业宏观分析和解决方案研究。

学者：AI 改善五大防御核心功能和助力七个攻击阶段

编者按：美国哥伦比亚大学国际与公共事务学院高级研究学者贾森·希利撰文，根据美国国家标准与技术研究所（NIST）的网络安全框架和洛克希德·马丁公司的“网络杀伤链”对人工智能对网络防御和网络攻击的促进作用进行了全面分析。

文章称，人工智能既可以成为网络防御的“力量倍增器”，也可以提升攻击者长期以来在网络空间中拥有的系统性优势。在网络防御方面，人工智能可以改善 NIST 网络安全框架提出的五个核心功能，包括识别、保护、检测、响应和恢复。在网络攻击方面，人工智能可以促进洛克希德·马丁公司“入侵杀伤链”提出的七个攻击阶段，包括侦察、武器化、投送、利用、安装、命令与控制，以及针对目标的行动。

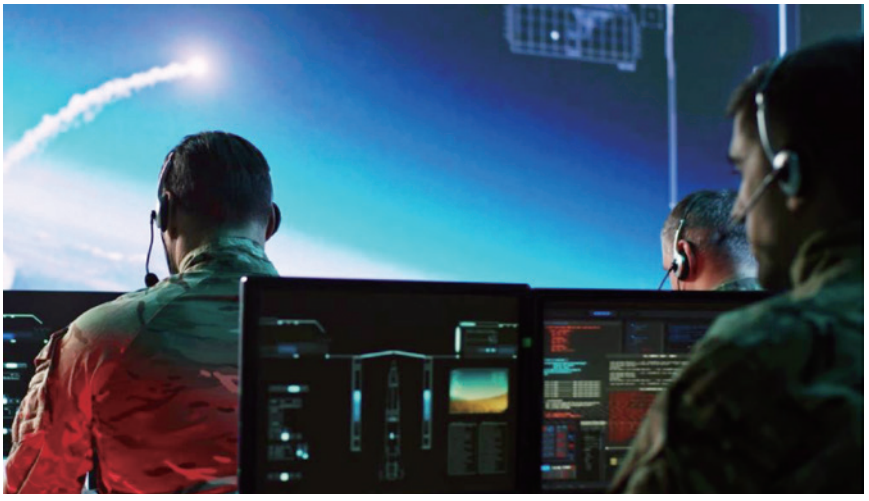
为了使网络空间更具防御性，创新不仅必须加强防御，还必须为防御者提供相对于攻击者的持续优势。

人工智能有潜力改变防守者的游戏规则。正如德勤最近的《网络人工智能：真正的防御》报告所述，“人工智能可以成为力量倍增器，使安全团队不仅能够比网络攻击者的行动更快地做出反应，而且能够预测这些行动并提前采取行动”。

然而，如果我们换个角度来看，这一点也同样正确：人工智能可以使网络攻击者的行动速度快于防御者的反应速度。

即使是最好的防御进步也会很快被攻击者的更大飞跃所超越，而攻击者长期以来在网络空间中拥有系统性优势。正如安全专家丹·格尔在 2014 年所说，“无论是在检测、控制还是预防方面，我们都在创造个人最好的成绩，但对手却一直在创造世界纪录。”最令人沮丧的是，许多有希望的防御措施——例如破解密码或扫描网络漏洞的“进攻性安全”——最终对攻击者的推动力超过了防御者。

为了让人工智能避免这种命运，防御者及那些资助新研究和创新的人必须记住，人工智能并不是“一根能带来持久无懈可击的魔杖”。为了让防御者赢得人工智能网络安全军备竞



赛，必须不断更新投资并有针对性地进行投资，以领先于威胁行为者自己对人工智能的创新使用。

很难评估人工智能会在进攻还是防御中提供更多帮助，因为每一方都是独一无二的。但可以使用两个广泛使用的框架来澄清这种“风马牛不相及”的比较。

美国国家标准与技术研究所（NIST）的网络安全框架可用于凸显人工智能帮助防御的多种方式，而洛克希德·马丁公司开发的网络杀伤链框架，也可以为攻击者使用人工智能做同样的事情。

这种更加结构化的方法可以帮助技术专家和政策制定者确定投资目标，并确保人工智能不会重蹈许多其他技术的覆辙，即推动防御者但也会加剧进攻。

一、人工智能在国防方面的收益

美国国家标准与技术研究所（NIST）的框架是一个理想的架构，涵盖了人工智能可能帮助防御者的所有方式。表 1 虽然不是完整列表，但可作为介绍。

尽管这只是一个子集，但仍然有很大的收获，特别是如果人工智能可以大幅减少高技能防御者的数量。不幸的是，大多数其他收益与攻击者的相应收益直接匹配。

NIST 框架功能	人工智能可能从根本上改善防御的方式
识别	- 快速自动发现机构的设备和软件
	- 更轻松地绘制机构的供应链及其可能的漏洞和故障点
	- 快速、大规模地识别软件漏洞
保护	- 减少对训练有素的网络防御者的需求
	- 降低网络防御者所需的技能水平
	- 自动修补软件和相关依赖项
检测	- 通过大规模、快速地检查数据，快速发现企图入侵的行为，几乎不会出现误报警报
响应	- 通过快速扫描日志和其他行为，大大改进对对手活动的跟踪
	- 无论在何处发现攻击者，都会快速自动驱逐
	- 更快的逆向工程和反混淆，以了解恶意软件如何工作以更快地挫败和归因
	- 用于人工跟踪的误报警报大幅减少
恢复	- 自动重建遭渗透的基础设施并以最短的停机时间恢复丢失的数据

表 1：使用 NIST 框架对防御者的人工智能优势进行分类

二、人工智能在进攻中的收益

虽然 NIST 框架是正确的防御工具，但洛克希德·马丁公司的网络杀伤链是一个更好的框架，用于评估人工智能如何促进军备竞赛的攻击方，这一想法是由美国计算机科学家凯瑟琳·费舍尔早些时候提出的。（MITRE ATT&CK 是另一个以犯罪为主题的框

网络杀伤链框架阶段	人工智能可能从根本上改善进攻的方式
侦察	- 自动查找、购买和使用被泄露和被盗的凭证
	- 自动排序以查找具有特定漏洞（广泛）的所有目标或有关确切目标的信息（深层；如详细说明硬编码密码的晦涩帖子）
	- 自动识别可能影响主要目标的供应链或其他第三方关系
	- 加快访问代理识别和聚合被盗凭证的规模和速度
武器化	- 快速、大规模地自动发现软件漏洞并编写概念验证漏洞
	- 大幅改善混淆，阻碍逆向工程和归因
	- 自动编写优质的网络钓鱼电子邮件，如通过阅读高管的大量信件并模仿他们的风格
	- 创建深度造假音频和视频来冒充高级管理人员以欺骗员工
投送、利用和安装	- 与许多机构的防御者进行实际的并行交互，说服他们安装恶意软件或执行攻击者的命令
	- 生成虚假攻击流量以分散防御者的注意力
命令与控制	- 更快的突破：自动权限升级和横向移动
	- 自动编排大量遭渗透的机器
	- 植入的恶意软件能够独立行动，无需与人工处理人员沟通以获取指示
针对目标的行动	- 以不易察觉的模式自动秘密泄露数据
	- 自动处理以识别、转换和汇总满足指定收集要求的数据

表 2：使用网络杀伤链框架对攻击者的人工智能优势进行分类

架，可能更好，但比一篇短文中可以轻松检查的要复杂得多。）

同样，尽管这可能只是人工智能协助犯罪的众多方式中的一个子集，但它展示了其可以带来的优势，特别是当这些类别组合在一起时。

三、分析和后续步骤

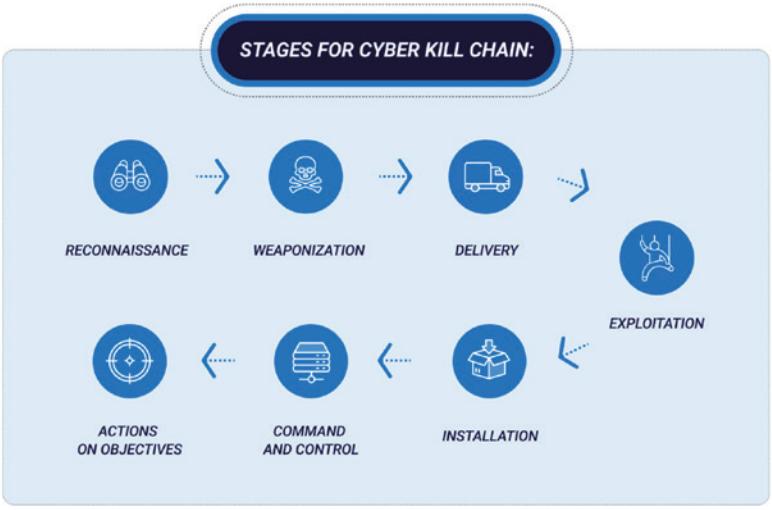
不幸的是，通用技术历来对攻击有利，因为防御者分散在组织内部和组织间，而攻击者则集中。为了充分发挥其作用，防御性创新通常需要在数千个组织（有时是数十亿人）中实施，而目标明确的攻击者群体可以更敏捷地整合进攻性创新。

这就是为什么人工智能对防御的最大帮助可能是减少所需的网络防御者的数量和他们所需的技能水平的原因之一。

仅美国就需要数十万额外的网络安全人员，而这些职位不太可能被填补。被雇佣的人需要花费数年时间来培养对抗高级攻击者所需的技能。此外，人类还要努力应对复杂而分散的任务，如大规模防御。

随着越来越多的机构将其计算和网络任务转移到云端，主要服务提供商将处于有利地位，可以集中人工智能驱动的防御。人工智能的规模可能会彻底改变防御，不仅对少数能买得起先进工具的人来说，而且对互联网上的每个人来说都是如此。

未来不是写在石头上的，而是写在代码里的。现在，明智的政策和投资可以发挥重大作用，使人工智能军备竞赛的平衡向防御倾斜。例如，负责开发军用技术的美国国防高级研究计划局（DARPA）正在进行变革性投资，显然是从经验中吸取了教训。



2016 年，DARPA 主办了“网络挑战赛”（Cyber Grand Challenge）的最后一轮比赛，旨在创建“有史以来开发的一些最复杂的自动漏洞搜寻系统”。但这些计算机既可以进攻也可以防守。为了获胜，他们“需要利用对手软件中的漏洞”并对其进行攻击。自主进攻系统可能是军队的一项自然投资，但不幸的是会增强进攻的优势。

DARPA 的新实验“人工智能网络挑战赛”（AI Cyber Challenge）纯粹是防御性的（没有进攻性的“夺旗”成分）。“利用人工智能的进步来开发，能够自动保护日常生活关键代码安全的系统”。这项 DARPA 挑战赛的奖金近 2000 万美元，并得到人工智能领先公司（Anthropic、Google、Microsoft 和 OpenAI）的支持，可能会彻底改变软件的安全性。

这两个挑战完美地概括了这一现实：技术专家和政策制定者需要开展投资，以确保防御性 AI 能够更快地发现漏洞、修补漏洞及其相关问题，而不是落后于进攻性 AI 发现、武器化和利用漏洞的速度。

预计 2021 年至 2025 年间，全球用于网络安全的人工智能支出将增加 190 亿美元，最终让防御方获得相对于进攻方优势的机会看起来非常光明。

关于作者

贾森·希利

美国哥伦比亚大学国际与公共事务学院高级研究学者。1998 年帮助创建了世界上首个网络司令部——美国计算机网络防御联合特遣部队，后来曾担任白宫网络政策主管。

极牛技术社群

网络安全技术社群

Cyber Security Technology Community

PLATFORM

网络安全技术社群

极牛网旗下面向网络安全工程师的社群平台，汇聚行业中网安工程师的知识和社交平台，围绕『技术』『管理』和『圈子』三个核心能力，将前沿的技术内容、技术管理成长感悟、技术圈子社群平台等向优质的网安工程师开放，每月定期组织社群会员线下活动，强调内容分享的体系化和活动内容的多样性，为中国网络安全工程师打造专属的全方位综合赋能的社群平台。

技术

聚焦前沿技术、热点技术、难点技术，提升网安在企业架构中的决策能力和迭代能力。

管理

专注在帮助网安工程师在职业发展中，建立体系化的管理知识和技术管理转型路径。

圈子

建立精准的技术方向圈子，针对不同的职业发展阶段，组成技术成长小组一起结伴同行。

极牛技术站

以沙龙、峰会、圆桌论坛等深度学习的形式进行，满足知识获取与社交需求。

管理加油站

面向管理者的闭门活动，前瞻性的思维和观点，成熟的管理模式与领导模型。

极牛知识变现

通过帮助工程师打造个人品牌，从出书、录课、企培等形式帮助知识变现。

极牛训练营

面向社群成员的线下课程，技术架构、团队管理、商业模式、塑造个人品牌等。

极牛众星计划

社群中优秀的工程师将会被受邀签约极牛众星计划，平台辅助进行品牌孵化。

更多内容敬请期待

华云信安

以人才梯队为依托，以自主创新为核心，深耕网络空间安全产业

华云信安(深圳)科技有限公司（简称 华云信安™ ）成立于2016年，是一家深耕于网络空间安全领域，拥有自主研发能力及核心知识产权，提供网络安全解决方案与技术服务的高新技术企业。华云信安™总部位于深圳，在广州、上海、武汉设有分支机构，公司核心团队来自奇安信、网易、华为、绿盟、思科等国际知名科技企业，具有深厚的网络安全技术实力和管理经验。

华云信安™目前拥有数十项计算机软件著作权和十余款自主研发产品，具备“风险评估类”和“安全工程类”两项信息安全服务资质，通过ISO9001质量管理体系认证，现为深圳市信息安全行业协会理事单位和深圳市信息网络安全管理协会会员单位。

华云信安™凭借创新的网络安全产品体系、深厚的网络安全服务能力以及丰富的服务经验，为互联网、金融、能源、制造、交通、医疗、政务等领域的广大客户，提供专业优质的网络安全产品、网络安全解决方案和网络安全技术服务。

网络犯罪研究中心

华云信安·网络犯罪研究中心，是专注于打击网络犯罪的安全服务部门，致力于打击涉网新型犯罪领域的安全技术研究与产品研发，包括涉网犯罪案件技术支撑、网络诈骗案件技术支撑、涉众型经济犯罪案件技术支撑等，以攻防实验室和极牛技术社群组成创新型的安全研究团队，为国家及各省市公安机关提供高效专业的打击涉网犯罪情报分析服务和实战解决方案。

极牛攻防实验室

华云信安·极牛攻防实验室，由内部成员及外部知名技术专家团队组成，致力于最前沿网络安全技术的研究和调研，以指导技术研发路径和产品发展方向。其职责除开展传统的网络安全技术研究外，还跟踪国内外网络安全技术趋势并进行相关技术研究。现已协助国家互联网应急中心(CNcert)发现并修复数百个安全漏洞，获得数十张高危原创漏洞证明证书。

涉网犯罪对抗　护网重保服务　安全咨询服务　安全保障服务　安全培训服务

丰富的解决方案

我们提供卓越、丰富的信息安全产品及各行业解决方案、信息安全技术服务和基础架构咨询服务。

专业的团队服务

我们拥有资深、专业的服务团队，按需为用户提供多元化、多级别的咨询规划服务和技术实施服务。

众多的行业案例

我们拥有众多行业应用案例，包括制造、快消、航空、金融、物流、建筑、服务、互联网等行业。

深入的厂商合作

我们与多家国内外知名厂商建立战略合作关系，共同致力于优质解决方案及相关服务的推广落地。

全国范围的服务

我们公司总部在深圳，同时在上海、广州、武汉等设有分支机构，具有全国范围内的业务服务能力。

公众号 小程序 官网

网安观察

没有网络安全就没有国家安全

7436084028