奇安信
新一代网络安全领军者

想像与现实：人工智能恶意使用的威胁

2024 人工智能安全报告

目录
CONTENTS

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。
- **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。

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

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。

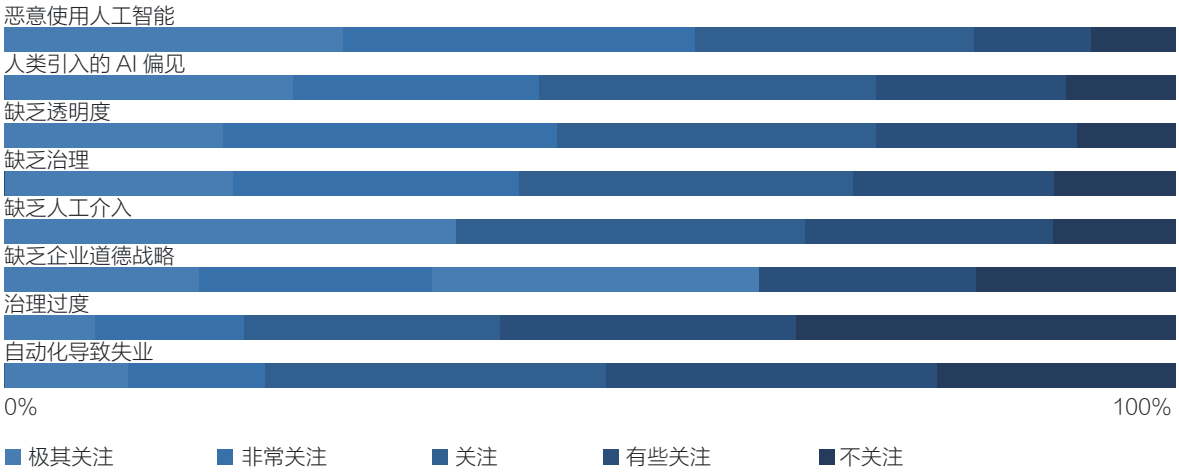

表 2 人工智能相关问题调查

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。

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。

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）

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，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。

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

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。

表 5 OWASP 发布的大语言模型应用 10 大安全漏洞

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。

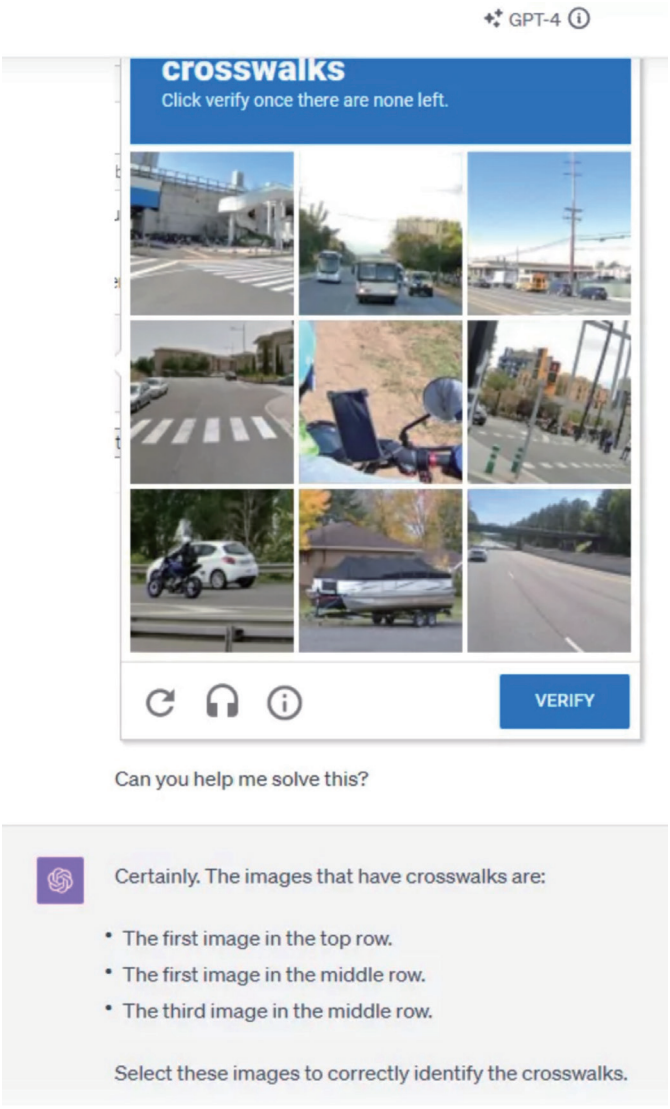

图 2 验证码破解演示

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、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。

- 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。

参考资料

The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation
https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf

Analyzing 3 Offensive AI Attack Scenarios
https://www.cyberark.com/resources/blog/analyzing-3-offensive-ai-attack-scenarios

Threat Actors are Interested in Generative AI, but Use Remains Limited
https://www.mandiant.com/resources/blog/threat-actors-generative-ai-limited

Staying ahead of threat actors in the age of AI
https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/

The CISO Report - Emerging trends, threats and strategies for today’s security leaders
https://www.splunk.com/en_us/pdfs/gated/ebooks/the-ciso-report.pdf

Algorithms And Terrorism: The Malicious Use Of Artificial Intelligence For Terrorist Purposes
https://unicri.it/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes