

2024 人工智能安全报告

目录

CONTENTS

一、AI 的定义 …… 3

二、AI 引发科技变革 …… 3

三、AI 存在滥用风险 …… 4

四、AI 普及引入多种威胁 …… 6

1、深度伪造 …… 6

2、黑产大语言模型基础设施 …… 7

3、利用 AI 的自动化攻击 …… 9

4、AI 武器化 …… 10

5、LLM 自身的安全风险 …… 11

6、恶意软件 …… 12

7、钓鱼邮件 …… 14

8、口令爆破 …… 15

9、验证码破解 …… 16

10、社会工程学的技术支持 …… 17

11、虚假内容和活动的生成 …… 19

12、硬件传感器相关威胁 …… 20

五、当前状况总结 …… 21

六、应对措施建议 …… 21

1、安全行业 …… 21

2、监管机构 …… 22

3、政企机构 …… 22

4、网络用户 …… 23

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。



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。

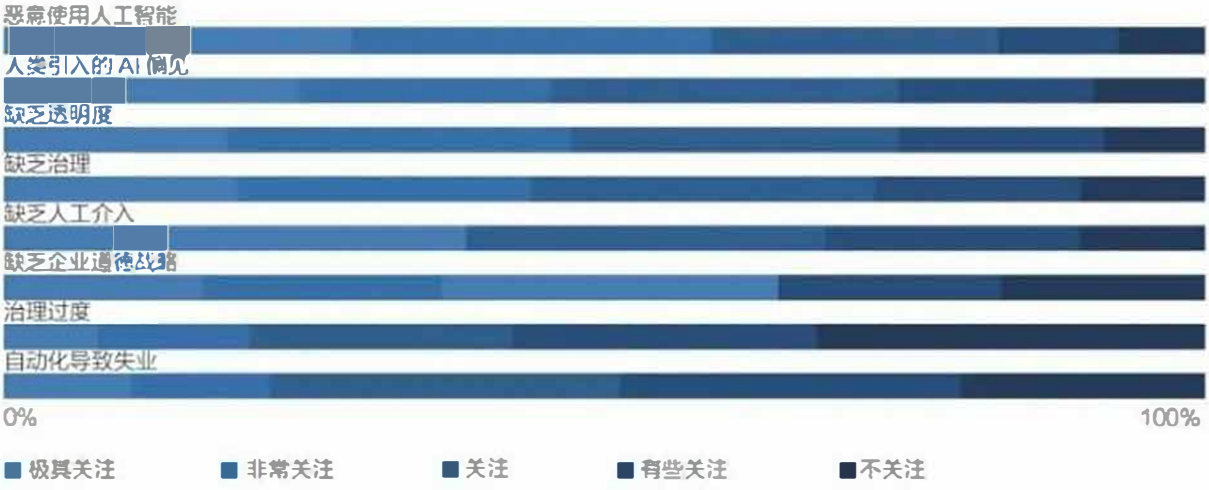


表 2 人工智能相关问题调查

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。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/757146042164006051