DOI 10.15302/J-SSCAE-2021.03.002

人工智能赋能网络攻击的安全威胁及应对策略

方滨兴[1]，时金桥[1]，王忠儒[2]，余伟强[3]

（1. 北京邮电大学网络空间安全学院，北京 100876；2. 中国网络空间研究院，北京 100010；
3. 北京丁牛科技有限公司，北京 100081）

摘要：人工智能（AI）在为社会进步带来显著推动效应的同时，也在促进网络空间安全领域的重大变革，研究 AI 和网络空间安全结合带来的安全问题具有迫切意义。本文采用自顶向下的分析方法，从加剧现实安全威胁、催生新型安全威胁两个角度分析了 AI 和网络空间安全结合带来的政治安全、经济安全、社会安全、国防安全等重大问题，提炼了自主化规模化的拒绝服务攻击、智能化高仿真的社会工程学攻击、智能化精准化的恶意代码攻击等新型威胁场景，总结了环境自适应隐蔽攻击、分布式自主协作攻击、自我演化攻击等未来发展趋势。为有效应对 AI 赋能网络攻击的安全威胁，建议从防范安全威胁、构建对等能力角度加强智能化网络攻防体系建设和能力升级；加强 AI 安全数据资产的共享利用，采取以数据为中心的 AI 网络攻防技术发展路径；加强对抗评估和测试验证，促进 AI 网络攻防技术尽快具备实用性。

关键词：人工智能；网络攻防；国家安全；自主协作；自我演化

中图分类号：TP393　　**文献标识码：**A

AI-Enabled Cyberspace Attacks: Security Risks and Countermeasures

Fang Binxing[1], Shi Jinqiao[1], Wang Zhongru[2], Yu Weiqiang[3]

(1. School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Chinese Academy of Cyberspace Studies, Beijing 100010, China; 3. Beijing DigApis Technology Co., Ltd., Beijing 100081, China)

Abstract: Artificial intelligence (AI) brings significant societal progress and it also revolutionizes the cybersecurity sector. Thus, studying the security problems induced by the deep fusion of AI and cyberspace security becomes significant. In this article, we systematically analyze the major national security issues induced by the fusion, involving political, economic, social, and national defense securities. These issues aggravate the existing security risks and trigger new threats. Moreover, new attack scenarios are analyzed, including autonomous and large-scale denial-of-service attacks, intelligent and disguised social engineering attacks, and intelligent and targeted malicious code attacks. Subsequently, future AI-enabled attack types such as situation-awareness covert attacks, distributed autonomous-collaboration attacks, and self-evolving attacks are explored. To effectively address the security threats of AI-enabled cyber attacks, we suggest that an intelligent network attack and defense system should be established and its capabilities upgraded to construct equivalent capabilities. Sharing of AI security data assets should be encouraged to develop a data-centered path for AI-enabled network attack and defense technologies. Furthermore, the AI-enabled network attack and defense technologies should be evaluated and verified through counterwork, enabling these technologies to be practically implemented.

Keywords: artificial intelligence (AI); cyber attack and defense; national security; autonomous collaboration; self-evolution

收稿日期：2021-02-28；**修回日期：**2021-04-23

通讯作者：时金桥，北京邮电大学网络空间安全学院教授，研究方向为网络空间安全；E-mail: shijinqiao@bupt.edu.cn

资助项目：中国工程院咨询项目“新一代人工智能安全与自主可控发展战略研究”(2019-ZD-01)

本刊网址：www.engineering.org.cn/ch/journal/sscae

一、前言

近年来，网络空间安全重大事件持续爆发，网络安全威胁全面泛化。斯诺登事件、乌克兰电网攻击事件、美国大选干预事件等表明，网络空间安全威胁覆盖了从物理基础设施、网络信息系统到社交媒体信息，对虚拟世界、物理世界的诸多方面构成威胁。网络空间安全已经成为非传统安全的重要组成部分。随着人工智能（AI）第三次浪潮的兴起，人工智能向诸多行业、领域不断渗透并交叉融合的趋势已经显现。人工智能因其智能化与自动化的识别及处理能力、强大的数据分析能力、可与网络空间安全技术及应用进行深度协同的特性，对网络空间安全的理论、技术、方法、应用产生重要影响，促进变革性进步。

人工智能与网络空间安全的交互融合，表现了“伴生”“赋能”两种效应 [1]。①网络空间安全在本质上是一种伴生学科，每一种新技术的出现都会引发伴生的安全问题；人工智能的伴生安全问题主要是内生安全问题、衍生安全问题，即由于人工智能自身在脆弱性、可预测性、可解释性等方面存在的安全隐患或问题，将自身安全问题转移或嫁接到人工智能应用上，使得人工智能系统自身或者应用人工智能技术的系统产生新的安全威胁；攻击者可利用对抗样本或数据投毒技术，自动化构造攻击样本，针对现有智能安全系统开展攻击，造成人脸识别、车牌识别等系统功能降级，甚至引导实施网络攻击、物理攻击 [2]。②人工智能在自身发展带来新网络空间安全威胁的同时，也从攻击、防御方面给传统网络空间安全提供了显著的赋能效应，如基于机器学习、深度搜索的人工智能方法能够提升网络攻击能力、自动检测网络安全防御方法、制定智能化的攻击策略；同样，人工智能可辅助网络空间安全从被动防御趋向主动防御，从而更快更好地识别威胁、缩短响应时间；网络空间的时空动态变化复杂，人工智能技术可关联分析日志、流量等不同渠道的数据，构造多维数据关联与智能分析模型的资产库、漏洞库、威胁库，实现对有效网络攻击的全面、准确、实时检测 [3]。

人工智能在攻防两方面的赋能效应，极大地推动了网络空间攻防对抗的发展，引发新的安全威胁，催生新的对抗手段。对于网络安全而言，人工智能是一把“双刃剑”；人工智能与网络空间安全深度结合，给经济、政治、社会、国防等领域带来新威胁、新问题的同时，也为各国网络空间安全发展提供了新机遇。本文系统分析人工智能在网络空间安全领域应用带来的安全问题，重点研究人工智能在网络攻击细分方向的赋能效应，总结提炼人工智能赋能网络攻击的新兴威胁场景、技术发展现状、未来发展趋势，以期为相关领域发展提供理论参考。

二、人工智能和网络空间安全深度结合带来的国家安全问题

（一）涉及的国家安全问题与威胁

1. 政治安全方面

随着网络技术的迅猛发展及广泛运用，网络政治作为一种新的政治形态呈现出来。公众可以借助多元化网络通道和途径，较为自由地进行政治表达和参与，影响政治过程，实现政治权利，但也可能引发各种政治安全问题。人工智能显著加剧了政治安全领域中的现实威胁。例如，在 2018 年 3 月曝光的“剑桥分析”事件中 [4]，商业智能公司利用脸书用户数据进行人物画像，自动推送信息以影响选民在美国大选、英国脱欧等政治事件中的投票倾向；该事件标志着数据智能从商业领域扩散至政治领域，使得单纯的网络数据安全问题上升为现实的政治安全隐患。

人工智能技术应用可引发使用数字自动化塑造政治影响等新兴安全威胁 [5]。例如，应用深度伪造技术生成逼真的捏造视频、音频，编造领导人丑闻，伪造新闻进行煽动；利用人工智能的自然语言生成技术，自动化构造信息并进行定制化的虚假宣传活动。这类具有数字自动化特征的深度伪造威胁，借助各类媒体传播虚假信息，具有极强的传播势能，可实现大规模、潜伏性的政治操纵和控制，将显著加剧网络空间政治安全威胁的影响力和对抗复杂性。

2. 经济社会安全方面

人工智能与网络安全深度结合将威胁和影响经济社会安全。随着相关行业、企业、公众对网络技术与应用依赖性的增加，与网络犯罪相关的经济社会风险也随之增长。《2018 年全球风险报告》认为

[6]，网络攻击问题已经成为仅次于极端天气、自然灾害之外的世界第三大威胁。利用人工智能、大数据技术，攻击者可以根据出生年月、电话、亲属、位置等关键个人信息，“量身定制”个性化的诱饵攻击，实现高度逼真的自动化社会工程攻击。

借助自动化、智能化工具，网络罪犯可以针对大规模目标开展高效、隐蔽的漏洞探测扫描，完成自动利用和攻击。人工智能技术驱动的智能化、自动化、规模化攻击，可为网络犯罪提供威胁更大、传统防御系统更难防范的技术手段与方法，所产生的破坏力也更强，严重威胁和影响了经济社会安全。

3. 国防安全方面

人工智能与网络攻防结合程度的不断加深，将极大改变传统信息作战的方式与手段。通过智能化的态势感知、情报分析、网络攻击与瘫痪，可形成军事先发优势并引发新型军备竞赛。在网络武器方面，人工智能为国家级高级可持续威胁攻击（APT）组织提供了新的工具与手段，针对关键信息基础设施实施渗透性、隐蔽性更强的网络攻击，严重影响其安全稳定运行。

2013 年，美国国防高级研究计划局（DARPA）发起的网络安全挑战赛（CGC）极大推动了自动化网络攻防技术的发展；基于人工智能的新型网络战武器将明显改变网络空间军事对抗格局，加速塑造不对称竞争优势。2017 年，美国成立算法战跨职能小组，加速将大数据、人工智能、机器学习整合到国防部项目，重点推动战场空间态势感知、自动化网络响应等技术研发。算法层面的突破、数据数量与质量的提升、计算能力的增长，为人工智能在国防领域的应用提供了巨大的想象空间，将构建新的战略威胁。

（二）主要国家的应对态势

1. 美国

美国凭借传统技术优势，积极谋求在人工智能技术方向的主导地位；将网络安全视为重要方面，高度重视人工智能在网络安全领域的研究与应用，争取建立网络攻防领域的战略优势。

2016 年，美国《为人工智能的未来做好准备》报告提出，相关机构的计划和战略应考虑人工智能、网络安全之间的相互影响；人工智能研究机构应确保人工智能技术自身及生态系统具备应对智能对手挑战、保持安全性和恢复力的优势；参与网络安全工作的机构应采用美国自有的人工智能技术来高效实现网络安全。同年发布的《人工智能、自动化与经济》报告认为，为有效应对人工智能自动化对经济的不利影响，应从网络防御、欺诈侦察的角度发展人工智能技术；典型应用有基于人工智能的机器学习系统辅助人类迅速回应网络攻击，人工智能高效解读数据并预防网络攻击。

2017 年，哈佛大学《人工智能与国家安全》报告指出，网络武器将更频繁地用于虚拟作战；机器学习在军事系统中应用，将带来新型漏洞并催生新型网络攻击手段；人工智能网络武器一旦被盗或者非法复制，将被恶意使用；不断进步的自动化将使失业问题、网络攻击问题更为严峻，进而影响政治稳定和国家安全。

2018 年，美国国际战略研究中心发布《人工智能与国家安全，人工智能生态系统的重要性》，报告认为，在网络安全或防御等领域，人类可能无法迅速作出反应，首先掌握人工智能应用的国家会有显著优势；在网络安全方面，人工智能技术可与僵尸网络配合，实施攻击并打垮防御。

2019 年，美国发布新版《国家人工智能研究与发展战略规划》，列出了算法对抗、数据中毒、模型反转等威胁人工智能安全的问题；要求在人工智能系统全生命周期考虑安全性问题，涵盖初始设计，数据 / 模型的构建、评估、验证、部署、操作、监视等环节。

2021 年 3 月，美国人工智能国家安全委员会发布建议报告，认为美国尚未做好防御人工智能赋能新兴威胁的准备；提出 2025 年实现军事人工智能战备状态的发展目标，建议成立技术竞争力委员会等组织机构，确保赢得竞争并增强防御能力。

2. 其他国家

2018 年，俄罗斯发布《人工智能在军事领域的发展现状以及应用前景》，明确将人工智能视为战略竞争的重要领域，推动人工智能元素与无人集群、无人自主系统反制、雷达预警系统的整合，支持国家军事能力提升。

2016 年，日本防卫省发布《中长期技术规划》，推动发展可快速处理海量情报数据的人工智能技术、能够应对网络攻击的广域分散情报通信系统技术，由此提升态势感知、情报共享、电子攻防、指

挥控制能力。

2018 年，印度发布《人工智能国家战略》，注重利用人工智能技术促进经济增长并提升社会包容性，寻求适合国情的人工智能规划部署。印度将利用人工智能技术开发武器、防御、监视系统，制定人工智能发展路线图；研究机器学习在军兵种、网络安全、核、生物资源等领域应用，以自主化武器、无人监视系统为代表。

三、人工智能赋能网络攻击的安全威胁场景与典型技术

（一）人工智能赋能网络攻击带来的新型威胁场景

1. 自主化、规模化的拒绝服务攻击威胁

近年来，随着物联网（IoT）的逐步普及、工控系统的广泛互联，直接暴露在网络空间的联网设备数量大幅增加。Mirai IoT 僵尸网络分布式拒绝服务攻击（DDoS）事件（2016 年）表明，攻击者正在利用多种手段控制海量 IoT 设备，将这些受感染的 IoT 设备组成僵尸网络，发动大规模 DDoS 攻击并可造成网络阻塞和瘫痪。除了呈现大规模攻击的典型特点之外，网络攻击者越发注重将人工智能技术应用于僵尸网络攻击，据此进化出智能化、自主化特征。

2018 年全球威胁态势预测 [7] 表明，人工智能技术未来将大量应用在类似的蜂群网络中，可使用数百万个互连的设备集群来同步识别并应对不同的攻击媒介，进而利用自我学习能力，以前所未有的规模对脆弱系统实施自主攻击。这种蜂巢僵尸集群可进行智能协同，根据群体情报自主决策采取行动，无需僵尸网络的控制端来发出命令；无中心的自主智能协同技术，使得僵尸网络规模可突破命令控制通道的限制而成倍增长，显著扩大了同时攻击多个目标的能力。人工智能赋能的规模化、自主化主动攻击，向传统的僵尸网络对抗提出了全新挑战，催生了新型网络空间安全威胁。

2. 智能化、高仿真的社会工程学攻击威胁

社会工程学利用人性弱点来获取有价值信息，作为攻击方法是一种欺骗的艺术。社会工程学网络攻击虽出现已久，但始终是较为有效的攻击手段；特别是鱼叉式网络钓鱼，因成效显著、传统安全性防御机制难以阻止而成为研究关注重点。随着人工智能应用的拓展，社会工程学攻击日益呈现智能化、高仿真特征。攻击者利用社交媒体等开放获取的个人隐私数据，自动学习并构造虚假信息，让受攻击目标不引起怀疑而自愿上钩。

在 2016 年美国黑帽大会上，网络安全公司 ZeroFOX 的安全研究员展示了一种带有侦察功能的社交网络自动钓鱼攻击方法 [8]；利用机器学习算法，通过网络大数据挖掘个人的出生年月、电话、亲属关系、位置等关键信息，自动生成定制化、高仿真的恶意网站 / 电子邮件 / 链接；模仿相关联系人的通信内容风格并骗取信任，从模仿真实联系人的地址发送出来，有效提升钓鱼攻击的有效率。利用人工智能技术，攻击者还可创建逼真的低成本伪造音频和视频，将网络钓鱼攻击空间从电子邮件扩展到其他通信域（如电话会议、视频会议），加剧了社会工程学攻击威胁。

3. 智能化、精准化的恶意代码威胁

随着人工智能技术的发展，攻击者倾向于针对恶意代码攻击链的各个攻击环节进行赋能，增强攻击的精准性，提升攻击的效率与成功率，有效突破网络安全防护体系，对防御方造成重大损失。在恶意代码生成构建方面，深度学习赋能恶意代码生成相较传统的恶意代码生成具有明显优势，可大幅提升恶意代码的免杀和生存能力。在恶意代码攻击释放过程中，攻击者可将深度学习模型作为实施攻击的核心组件之一，利用深度学习中神经网络分类器的分类功能，对攻击目标进行精准识别与打击。

在 2018 年美国黑帽大会上，国际商业机器公司（IBM）研究院展示了一种人工智能赋能的恶意代码 DeepLocker [9]，借助卷积神经网络（CNN）模型实现了对特定目标的精准定位与打击，验证了精准释放恶意代码威胁的技术可行性。目前，这类攻击手法已被攻击者应用于实际的高级持续性威胁攻击，一旦继续拓宽应用范围，将难以实现对抗防范；如果将之与网络攻击武器结合，有可能提升战斗力并造成严重威胁和破坏。

（二）人工智能赋能网络攻击的典型技术

1. 网络资产自动探测识别技术

网络资产探测识别指追踪、掌握网络资产情况的过程。从安全攻击的角度看，网络资产探测识别可用于渗透（或攻击）前的信息收集，了解目标网

络内主机的操作系统类型、开放端口以及所运行的应用程序类型与版本信息。准确掌握目标网络的安全状况，有助于选取高效的攻击方法。

在网络资产探测识别的人工智能应用方面，当前最具代表性的技术应用是基于机器学习的操作系统指纹识别技术。引入机器学习、深度学习等方法，进行操作系统指纹识别，可以较短的建模时间、较高的准确率，实现基于协议栈指纹被动操作系统的识别，提高未精确匹配指纹的识别率。

2. 智能社会工程学攻击技术

自动化社会工程学攻击技术指利用机器学习、神经网络等方法，实现钓鱼式攻击、电脑蠕虫传播、垃圾邮件散发等的完整攻击过程自动化。基于自然语言生成（NLG）的自动化网络钓鱼是一种典型攻击方法，攻击者利用深度学习分析文本内容，识别目标感兴趣的主题，生成目标可能响应的文本内容；常用于以电子邮件、社交网站作为攻击代码传输载体的新型网络钓鱼攻击。

在 2016 年第七届新西兰黑客大会上，意大利安全专家发布了一种自动化网络钓鱼工具 [10]，在在对澳大利亚公务员的调查测试中，成功欺骗了 40% 的参与人员。2019 年，有研究基于 NLG 技术构建了高级电子邮件伪装攻击生成引擎 [11]，评估实验表明，生成的伪装电子邮件具有更好的连贯性、更少的语法错误，是效果更优的网络钓鱼电子邮件攻击手段。

3. 智能恶意代码攻击技术

机器学习算法已经普遍应用于网络安全检测领域，然而相关检测系统容易受到对抗性攻击；攻击者可以构造“良性”样本，成功绕过机器学习分类器的识别。对抗机器学习在恶意代码中插入一部分对抗性样本，可绕过安全产品的检测；甚至根据安全产品的检测逻辑，自动化地在每次迭代中自发更改代码和签名形式，确保自动修改代码逃避反病毒产品检测且功能不受影响。2018 年，有研究利用深度强化学习网络提出了一种基于对抗样本生成的黑盒攻击方法，用于攻击静态的可执行文件（PE）杀毒引擎 [12]。这是第一个可以产生对抗性 PE 恶意代码的研究工作，模拟真实攻击的成功率达到 90%。随着人工智能在对抗机器学习领域的拓宽应用与进化，可以预见，基于生成对抗网络的逃逸攻击会成为对抗机器学习方面的重要方向和技术趋势。

此外，在传统恶意代码被发布后，攻击目标和意图往往是确定的，可通过逆向工程、网络监听等方法分析得知。在人工智能技术的助力下，恶意代码通过内嵌深度神经网络模型，可在代码开源的前提下依然确保攻击目标、攻击意图、高价值载荷的高度机密性，由此显著提升攻击的隐蔽性。此类攻击的代表性成果是 IBM 研究院的 DeepLocker 恶意代码。

4. 自动化漏洞挖掘与利用技术

自动化漏洞挖掘与利用指在无人工干预的基础上，自动化挖掘软件内部缺陷并利用该缺陷使软件实现非预期功能。2013 年，DARPA 发起了 CGC 项目，旨在实现漏洞挖掘、分析、利用、修复等环节的完全自动化，进而建立具备自动化攻击与防御能力的高性能网络推理系统。2014—2016 年，CGC 比赛在漏洞自动攻防方向进行了尝试，引起广泛关注。参赛团队建立自动攻击防御系统，实现无人干预条件下的自动寻找程序漏洞、自动生成漏洞利用程序攻击敌方、自动部署补丁程序抵御对手攻击的基本能力。国内自 2017 年起组织开展了类似的自动攻防比赛，促进了相关技术发展和新型网络安全系统构建。

四、人工智能赋能网络攻击的技术发展趋势

随着人工智能与网络安全的深度结合，人工智能赋能网络攻击与传统网络攻击在技术与手法上相比，将使过去劳动密集型、成本高昂的攻击手法开始彻底转型，朝着分布式、智能化、自动化方向发展，从而形成更为精准和快速的自动化攻击手法。相关技术发展趋势有以下三点。

一是利用人工智能学习环境特征，增强攻击的适应性与隐蔽性。在攻击目标的网络环境中，数据、行为等均具有一定的本地化特征。攻击者利用人工智能对目标网络中的数据、行为等特征等进行收集和建模，学习目标网络环境中正常的数据内容、传输频率、传递方法等环境特征；参考环境特征来选择合适的攻击手段，将攻击数据伪装成目标网络中具有正常特征的普通数据，将攻击行为伪装成目标网络中正常用户的网络行为；实现环境自适应的攻击行为、数据隐藏，提升攻击的隐蔽

性，增强攻击的适应性。

二是利用人工智能增强分布式协作效果，提高攻击的鲁棒性。攻击者引入分布式智能协同算法，将传统的由智能中心统一调度分布式攻击实体开展协作攻击，演化为无中心的分布式多智能攻击实体的自主协同和群体决策，从而提高多个分布式攻击节点之间的协作效率，降低对中心化协同调度的依赖性，减少攻击反制的风险，提升攻击的鲁棒性。

三是利用人工智能实现攻击方式的自我进化，提升攻击的有效性。攻击者利用人工智能分析不同攻击方式下的攻击效果及防御方的可能应对措施，进而针对防御方的弱点自动选择新的攻击机制，据此实现攻击方式的智能进化。例如，攻击者可将防御方入侵检测系统的结果作为反馈，采用人工智能技术对反馈数据进行收集和建模分析，建立攻击效果模型，动态调整合适攻击方式，规避入侵检测系统。

五、人工智能赋能网络攻击威胁的应对建议

（一）强化研究与应用，推动智能化网络攻防体系建设和能力升级

着眼人工智能赋能网络攻击的威胁和影响，从防范安全威胁、构建对等能力的视角着手，尽快开展重大关键技术研究。推动“产学研”机构以有效应对人工智能赋能攻击新型威胁场景为首要需求，从攻防两方面进行联合攻关，开展智能化威胁态势感知、自动化漏洞挖掘与利用、智能恶意代码等技术研究。加快人工智能技术在国家、重要行业关键信息基础设施安全防护方面的体系化应用，整体性完成智能化升级换代，大幅提升关键信息基础设施安全保障、网络安全态势感知、网络安全防御、网络威慑的能力水平。为管控人工智能带来的新型网络安全威胁，应加强相关法律法规建设，规范人工智能网络安全健康发展，延缓并阻止与特定威胁相关的活动。

（二）加强共享和利用，破解人工智能网络攻防技术体系建设的数据难题

人工智能训练数据集既是人工智能安全研究中最有价值的数字资产，又是关乎人工智能安全能力建设成功与否的战略资产。然而，目前人工智能安全训练数据缺乏安全、可控、可追溯的手段进行共享利用，这成为限制人工智能攻防技术快速发展的重要因素之一。建议以国家实验室等权威机构为依托，利用区块链等新型技术构建人工智能数据靶场，形成安全可信、激励机制合理的共享利用框架，促进人工智能数据资产的有效利用，落实以数据为中心的人工智能网络攻防技术发展路径。

（三）加强对抗和评估，促进人工智能网络攻防技术实用性发展

人工智能攻防属于持续对抗升级的技术，实际应用效果依赖对抗环境的全面性和真实性。然而由于科研条件尚不充分，现有人工智能攻防技术研究难以复现实际的攻防对抗环境，对人工智能自动化攻防技术从理论走向实际构成明显制约。

建议以国家实验室等权威机构为依托，构建人工智能攻防对抗靶场，通过权威评估、技术挑战赛、测试验证等形式，有效推动人工智能网络攻击、自动化漏洞发现与利用的效能评估和对抗分析，促进人工智能攻防技术加速朝着实用方向发展。

六、结语

网络空间安全威胁全面渗透虚拟世界和物理世界，给各国的政治、经济、社会和国防带来了巨大的安全风险和挑战。人工智能与网络空间安全威胁的深度结合，则进一步加剧现实安全威胁，催生新型安全威胁，给国家安全带来了更加严峻的挑战。人工智能赋能网络攻击，在大数据等关联技术的辅助下，使网络攻击愈发呈现出大规模、自动化、智能化等新的特点，必将带动和促进网络空间防御技术、手段、能力的进化与发展。

当前，在人工智能赋能网络攻防的发端之际，谁抢先找到人工智能与网络攻防在技术、数据、模型等层面的最佳结合点，抢先形成网络攻防领域的“技术差”“应用差”，谁就可能抢占网络空间对抗的技术制高点，从而形成对抗博弈优势，掌握网络空间主动权和威慑力。我国应加强人工智能在网络空间安全领域的战略应用，从防范新型威胁、积极应对挑战两个方面开展工作，着力解决人工智能在网络攻防领域应用中面临的数据、对抗、评估等实际问题，推动人工智能攻防尽快从理论研究

走向实际应用。识别人工智能带来的新型网络空间安全威胁，提升智能威胁感知应对能力，确保在人工智能变革的背景下有效维护国家网络空间主权，保障网络空间核心利益，为国家安全和发展保驾护航。

参考文献

[1] 方滨兴. 人工智能安全 [M]. 北京: 电子工业出版社, 2020.
Fang B X. Artificial intelligence safety and security [M]. Beijing: Publishing House of Electronics Industry, 2020.

[2] Gu Z Q, Hu W X, Zhang C J, et al. Gradient shielding: Towards understanding vulnerability of deep neural networks [J]. IEEE Transactions on Network Science and Engineering, DOI: 10.1109/TNSE.2020.2996738.

[3] Jia Y, Gu Z Q, Li A P, et al. MDATA: A new knowledge representation model – Theory, methods and applications [M]. Cham: Springer International Publishing, 2021.

[4] Loganzhu. 还原Facebook史上最大数据外泄事件始末 [EB/OL]. (2018-03-21)[2021-02-26]. https://stock.qq.com/a/20180321/004747.htm.
Loganzhu. Restore the story of the biggest data breach in Facebook's history [EB/OL]. (2018-03-21)[2021-02-26]. https://stock.qq.com/a/20180321/004747.htm.

[5] Brundage M, Avin S, Clark J, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation [EB/OL]. (2018-05-05)[2021-02-26]. https://maliciousaireport.com/.

[6] World Economic Forum. The global risks report 2018 [EB/OL]. (2018-01-17)[2021-02-26]. https://cn.weforum.org/reports/the-global-risks-report-2018.

[7] Fortinet. Fortiguard labs 2018 threat landscape predictions [EB/OL]. (2017-11-14)[2021-02-26]. https://www.fortinet.com/blog/business-and-technology/fortinet-fortiguard-2018-threat-landscape-predictions.html.

[8] Seymour J, Tully P. Weaponizingdata science for social engineering: Automated E2E spear phishing on Twitter [EB/OL]. (2016-05-05)[2021-02-26]. https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf.

[9] Kirat D, Jang J Y, Stoecklin M. DeepLocker-concealing targeted attacks with 人工智能 locksmithing [C]. Las Vegas: Proceedings of Black Hat, 2018.

[10] Antisnatchor. Practical phishing automation with phishlulz [C]. Wellington: Proceedings of the Kiwicon X, 2016.

[11] Orru M, Muraena T G. The unexpected phish [C]. Amsterdam: Proceedings of the Hack in the Box Security Conference, 2019.

[12] Anderson H S, Kharkar A, Filar B, et al. Learning to evade static PE machine learning malware models via reinforcement learning [EB/OL]. (2018-01-20)[2021-02-26]. https://arxiv.org/abs/1801.08917.