DOI 10.15302/J-SSCAE-2021.03.003

基于人工智能的网络空间安全防御战略研究

贾焰[1]，方滨兴[2]，李爱平[3]，顾钊铨[2]

（1. 哈尔滨工业大学（深圳）计算机科学与技术学院，深圳 518055；2. 广州大学网络空间先进技术研究院，广州 510006；3. 国防科技大学计算机学院，长沙 410073）

摘要：网络空间是继陆、海、空、天之后的第五大活动空间，维护网络空间安全是事关国家安全、国家主权和人民群众合法权益的重大问题。随着人工智能技术的飞速发展和在各领域的应用，网络空间安全面临着新的挑战。本文分析了人工智能时代网络空间安全面临的新风险，包括网络攻击越来越智能化，大规模网络攻击越来越频繁，网络攻击的隐蔽性越来越高，网络攻击的对抗博弈越来越强，重要数据越来越容易被窃取等；介绍了人工智能技术在处理海量数据、多源异构数据、实时动态数据时具有显著的优势，能大幅度提升网络空间防御能力；基于人工智能的网络空间防御关键问题及技术，重点分析了网络安全知识大脑的构建及网络攻击研判，并从构建动态可扩展的网络安全知识大脑，推动有效网络攻击的智能化检测，评估人工智能技术的安全性三个方面提出了针对性的发展对策和建议。

关键词：人工智能；网络空间安全；网络攻击；网络防御

中图分类号：TP393　　**文献标识码：**A

Artificial Intelligence Enabled Cyberspace Security Defense

Jia Yan[1], Fang Binxing[2], Li Aiping[3], Gu Zhaoquan[2]

(1. College of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China;
2. Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China;
3. College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: Cyberspace is regarded as the fifth largest activity space following land, sea, air, and space. Protecting cyberspace security is a major issue related to national security, national sovereignty, and the legitimate rights and interests of the people. With the rapid development of artificial intelligence (AI) technology and its application in various fields, cyberspace security has been facing new challenges. This study analyzes the new risks of cyberspace security in the era of AI, such as more intelligent network attacks, more frequent large-scale network attacks, higher concealment of network attacks, stronger confrontation game of network attacks, and easier exposure to stealing of important data. AI technology has significant advantages in dealing with massive data, multi-source heterogeneous data, and real-time dynamic data, which can significantly improve the defense capability of cyberspace. This study introduces some key problems and technologies of AI-enabled cyberspace security defense, particularly the construction of a cyberspace security knowledge brain and the detection of network attacks. Furthermore, we propose the corresponding countermeasures and suggestions from three aspects: the construction of a dynamic and scalable network security knowledge brain, the promotion of intelligent detection against network attacks, and the evaluation of AI technologies' security.

Keywords: artificial intelligence (AI); cyberspace security; cyber attacks; cyber defense

收稿日期：2021-01-08；**修回日期：**2021-02-22

通讯作者：顾钊铨，广州大学网络空间先进技术研究院教授，研究方向为网络空间安全；E-mail: zqgu@gzhu.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

网络空间是构建在信息通信技术基础设施之上的人造空间，用以支撑人们在该空间开展各类与信息通信技术相关的活动。网络空间已经成为继陆、海、空、天之后的第五大活动空间，网络空间安全是国家安全的重要组成部分，网络空间频频发生的安全事件，已经严重影响了社会稳定和人民生命财产的安全，维护网络空间安全已成为事关国家安全、国家主权和人民群众合法权益的重大问题。

近年来，随着大数据、云计算、人工智能等新型信息技术的飞速发展，网络空间安全领域的一些难题得到解决，但是新技术也逐步被不法分子利用，造成了网络空间中新的安全问题与挑战。以人工智能技术为例，人工智能技术既能用于提升网络空间安全能力，也能引发新的安全问题，即人工智能的“赋能效应”和“伴生效应”[1]。

传统网络安全领域检测网络攻击主要依靠规则、模式匹配等方式，从流量数据、日志数据中检测符合一定规则和模式的数据。然而，随着网络安全数据量的飞速增长，基于规则、模式匹配的检测方式效果差，很难发现复杂的攻击事件 [2]。人工智能技术可以从大量的数据中自动学习，挖掘数据中蕴含的规律，近年来已被用来解决网络安全问题，例如，麻省理工学院计算机科学与人工智能实验室和创业公司 PatternEx 共同开发的 AI^2 网络安全入侵检测平台，可以准确预测 85% 以上的网络攻击 [3]；为了有效检测恶意 Powershell，通过将 Powershell 转换为文本数据，构造卷积神经网络大大提升了检测率 [4]。

人工智能技术同样可能被恶意利用，引发网络空间中新的安全问题，即人工智能技术的伴生效应。人工智能技术应用于漏洞挖掘时，能发现系统中存在的多个漏洞，从而导致系统更容易被攻击；人工智能应用于大规模网络攻击时，攻击者可以自适应生成攻击程序，通过大量的客户端实现智能化、自动化的网络攻击；人工智能应用于复杂的网络攻击时，攻击者可以隐藏攻击行为、攻击路径等，从而使得防御者更难发现和检测此类攻击；人工智能技术也可应用于对网络攻防的博弈，人工智能技术自身存在脆弱性，攻击者可以攻击部署在系统中的智能模型，造成防御模型失效；人工智能技术还可能被用来窃取用户的重要数据，通过对系统中各类数据的深度挖掘，关联分析、还原出用户的重要数据，从而引发更为严重的安全问题。

网络空间安全是一个攻防博弈的过程。当新的安全问题出现时，防御者需要针对性地予以解决。人工智能时代，由于智能化的提升，网络空间安全面临着更加严峻的风险与挑战，而人工智能技术也正是防御者维护网络空间安全的武器。如何基于人工智能技术提升网络空间安全的主动防御能力、应对人工智能时代网络空间安全的新风险与新挑战已成为一个十分迫切的问题。

本文在中国工程院“新一代人工智能安全与自主可控发展战略研究”重大咨询项目的支持下，分析了人工智能时代网络空间安全面临的风险与挑战；介绍了人工智能技术的优势与特点，以及其如何应用于网络空间安全；同时介绍了基于人工智能的网络空间安全防御关键问题及技术；最后针对网络空间防御提出了发展对策与建议。

二、人工智能时代网络空间安全面临的风险与挑战

（一）网络攻击越来越智能化

人工智能技术使得网络漏洞更容易被挖掘，各种恶意软件可以更便捷地生成和应用，从而造成网络空间面临更严峻的安全威胁。

美国国家漏洞数据库（NVD）、国家信息安全漏洞库（CNNVD）等历年来披露的漏洞越来越多，涵盖的设备、软件系统等也越来越多。人工智能技术的发展为漏洞的挖掘和利用提供了便利。模糊测试是一种自动化或半自动化的软件测试技术，构造随机、非预期的畸形数据，测试并监控程序执行过程中可能产生的异常及漏洞的可利用性。此类模糊测试技术又可以分为白盒、黑盒、灰盒模糊测试等，能高效地挖掘和利用程序漏洞 [5]。漏洞自动利用一般而言包括信息提取、漏洞识别、路径发现、状态求解及代码生成 [6]，通过从可执行文件、源码等输入数据中提取有用的信息，利用路径发现与状态求解获取利用案例，并生成漏洞利用的程序或数据，实现漏洞的自动化利用。

美国国防部高级研究计划局（DARPA）于 2013 年发起全球性网络安全挑战赛，旨在推进自

动化网络防御技术的发展，即实时识别系统缺陷和漏洞，并能自动完成补丁和系统防御等。2016 年，美国在拉斯维加斯举办了信息安全界的顶级赛事 Defcon CTF [7]，1 支名为 Mayhem 的机器夺旗赛（CTF）战队与另外 14 支人类顶尖的 CTF 战队进行角逐，机器战队一度超过两支人类战队，开创了自动化攻防的新局面。自动化攻防是网络空间安全面临的新挑战，自动化的网络攻击手段将加剧网络空间的安全威胁与挑战。

（二）大规模网络攻击越来越频繁

在人工智能时代，大规模的网络攻击越来越频繁。大规模网络攻击的形式主要包括拒绝服务攻击（DDoS）、域名解析服务器（DNS）劫持等。大规模网络攻击的目标也从传统的网络系统，延伸到物联网、工业设备、智能家居、无人驾驶系统等。

2016 年 10 月，美国多个公司的服务器遭到大规模分布式拒绝服务攻击，据报道，此次攻击涉及数百万互联网地址和恶意软件的大规模攻击，而这些攻击的来源主要是被 Mirai 僵尸网络感染的连网设备。近年来，此类大规模僵尸网络驱动的分布式 DDoS 可以利用数以万计的被感染的物联网设备，通过这些设备向受害网站发送大量流量，实现攻击。2018 年，美国曾组织专家讨论了针对无人驾驶汽车的攻击，其中包括大规模网络攻击可能造成的危害，并建议提前进行规划演练。人工智能技术还可生成可扩展攻击的智能僵尸网络。美国飞塔（Fortinet）公司在其发布的 2018 年全球威胁态势预测 [8] 中表示，人工智能技术未来将被大量应用在蜂群巢网络（Hivenet）和机器人集群（Swarmbots）中，能够利用大规模互连的设备或机器人集群同时识别和应对不同的攻击媒介，并利用自我学习能力实现前所未有的大规模自主攻击。

人工智能技术使得网络攻击的成本越来越低，可利用的攻击武器和资源越来越多，从而导致大规模的网络攻击越发频繁。

（三）网络攻击的隐蔽性越来越高

传统的网络攻击行为一般会在系统中留下痕迹，容易被追溯；攻击行为的目标和意图比较明确，容易被发现。人工智能时代，利用智能化技术可以对复杂的攻击行为进行隐藏，如通过不同的终端设备实施攻击，在不同的时间发动攻击等。

传统的恶意代码、恶意程序在发布以后，这些代码和程序的攻击目标、攻击意图往往是确定的，作为网络空间中的防御者，可以通过逆向工程、网络监听等方式分析得知攻击的目标和意图。然而，在人工智能技术的助力下，恶意代码、恶意程序可以通过内嵌深度神经网络模型，实现在代码开源的前提下，依然确保攻击目标、攻击意图、高价值载荷三者的高度机密性，从而大幅度地提升了攻击行为的隐蔽性。2018 年 8 月，国际商业机器公司（IBM）研究院在 Black Hat USA 2018 大会上展示了 AI-Powered Malware—DeepLocker [9]，借助人工智能技术实现了目标识别精准性和攻击载荷机密性，能有效对抗人工逆向分析。

高级持续性威胁（APT）攻击是一种集合了多种攻击方式的复杂攻击。攻击者往往会花很长时间对目标网络进行观察，针对性地搜集信息，并有针对性地发动攻击。这些攻击行为可以分布在很多设备上，不同攻击行为之间也可能存在很大的时间间隔，结合人工智能技术可以对攻击行为进行更好的设计和组合，从而躲避防御者的检测，保持攻击行为的高隐蔽性。

（四）网络攻击的对抗博弈越来越强

网络空间安全是一个攻防博弈的过程。人工智能技术在处理海量、多源异构数据方面具有巨大的优势，攻击者会使用人工智能技术构造规模更大、隐蔽性更强、后果更严重的攻击，而防御者则会利用人工智能技术去提升网络攻击检测的准确率，提高网络攻击检测效率，降低网络攻击误报率等。在这个过程中，人工智能技术促使网络空间的攻防博弈程度愈演愈烈。

在恶意软件识别方面，基于生成对抗网络（GAN）的 MalGAN 算法可以使用一个替身检测器来适配黑盒恶意软件检测系统，该算法生成的恶意代码能够绕过基于机器学习的检测系统 [9]。类似的，为了躲避 PDF 恶意软件检测器，基于遗传算法的对抗机器学习方法可以在保留自身恶意行为的前提下，绕过机器学习分类器的识别，让恶意检测器将其识别为良性样本 [10]。

此外，由于人工智能技术自身存在脆弱性，例如，图像识别神经网络容易被生成的和原样本高度

相似的对抗样本迷惑，造成错误识别 [11]；推荐系统容易被个别关键词影响，造成推荐结果被人为干预 [12]。当缺乏解释性的人工智能技术用于网络攻击或防御时，另一方则可利用模型自身的脆弱性发动防御或攻击，引发新一轮的网络攻防博弈。

（五）重要数据越来越容易被窃取或破坏

数据是一项重要的资源和资产，大型企业特别是互联网企业拥有着大量的用户数据，这些企业的系统一旦被攻击，很容易造成大规模的数据被窃取或破坏。除了互联网企业，很多传统企业也拥有重要的数据，而传统企业的安全意识不足，攻击者更容易通过技术手段从中窃取用户和企业的重要数据。人工智能技术则加剧了该情况的出现，攻击者利用人工智能技术能更加容易地窃取重要数据，或者破坏企业的核心数据。

在数据发布过程中，用户的数据很有可能由于匿名保护等程度不够，攻击者通过多种攻击方式可以获取到用户的数据，如偏斜攻击等。成员推断攻击可以用于获取训练数据集的关键信息，攻击者可以判断某条信息是否存在于目标模型的训练数据集中，从而实现针对重要数据的窃取。攻击者通过训练出多个模仿目标模型的影子模型，利用影子模型的识别结果去判断目标模型的训练集中是否包含某敏感数据 [13]。类似的，模型倒推攻击可以通过模型的输出反推训练集中某条目标数据的部分或全部属性值，攻击者在仅获得模型参数的情况下，就能够使用基于生成对抗网络的方式实现模型反演，重建出训练数据，造成数据被窃取 [14]。

三、人工智能在网络空间安全中应用的优势与特点

人工智能发展迅速，随着海量数据的积累、算法算力的大幅度提升，人工智能已成为目前最为热门的研究方向之一。

人工智能主要包括三大学术流派：符号主义、连接主义、行为主义。其中符号主义是一种基于问题、逻辑和搜索的高级符号处理体系，通过将信息和行为抽象到基于符号规则的系统中，并利用计算机逻辑推理模拟人类的抽象思维，代表性的成果包括专家系统、知识图谱 [15]、多维数据关联与智能分析（MDATA）模型 [16] 等。连接主义采用基于网络连接机制和学习算法进行建模，典型的成果包括感知机、深度神经网络 [17] 等。行为主义则认为智能是通过对环境反馈的自主感知做出相应的行为。

网络空间安全相关的数据体量大、数据种类多、数据增长快，传统的分析技术在处理此类数据时效率低、准确率低。人工智能在处理海量数据、多源数据、动态数据等方面具有显著的优势，能助力于网络空间安全，提升网络防御能力。

（一）海量数据的快速处理能力

网络空间安全相关的数据体量大，例如系统中保存的日志数据、网络流量数据等，处理如此海量的数据既需要庞大的算力支撑，也需要能处理如此海量数据的智能算法。由于人工智能技术能从海量数据中学习数据的特征，根据特征再对数据进行分类、聚类等处理，能大幅度提升效率和准确度。

以恶意代码检测为例，可以通过提取恶意代码的静态特征和动态特征进行智能化检测，其中静态特征包括文件散列（Hash）、签名特征、应用程序编程接口（API）函数调用序列、字符串特征等，动态特征则包括中央处理器（CPU）利用率、内存消耗、网络行为特征、主机驻留行为等，通过自动化提取或者经过特征工程提取的各类特征，可利用深度学习或机器学习方法，如卷积神经网络等，自动对可疑的恶意代码进行判定。基于人工智能技术的恶意代码检测，相比静态检测、动态检测、启发式检测和虚拟机检测等技术，能大幅度提升检测效率，并提高检测的准确率。

（二）多源异构数据的高效关联能力

网络安全相关的数据种类繁多、来源广泛，如通过传感器、网络爬虫、日志收集系统等能采集到不同类型的数据，从来源上数据类型可以分为环境业务数据、网络层数据、日志层数据、告警数据等类别，综合不同来源的异构数据进行综合分析能提升网络空间主动防御能力。

以网络安全态势感知为例，防御者需要对网络系统的资产状态进行全方位的掌握，因此需要获取各种来源的信息，包括资产信息、漏洞信息、攻击行为信息等，而这些信息往往又是通过流量数

据、日志数据等不同数据进行综合分析得到的结果。MDATA 模型有助于实现全方位的网络安全态势感知 [16]。该模型构建了不同类型的网络安全知识库，包括资产知识库、漏洞知识库、威胁知识库，其中资产知识库主要包括系统中的各类软硬件资产及运行状态信息等，漏洞知识库包括各种类型的漏洞，威胁知识库包括针对系统的各类攻击行为，资产知识库和漏洞知识库进行关联，可及时发现系统中的漏洞；漏洞知识库和威胁知识库关联，可发现攻击路径、攻击方法等，及时制定相应的防御策略；资产知识库和威胁知识库关联，可发现攻击者的攻击目标等，增加资产的保护力度等。基于人工智能技术的网络安全态势感知有助于实现网络空间主动防御，大幅度提升网络系统防御能力。

（三）动态数据的实时在线处理能力

网络空间安全相关的数据增长速度快，时效性要求高。从数据增长速度上来看，每天都会产生很多新的流量数据、日志数据、告警数据等，如何对这些新产生的动态数据进行分析是一个十分迫切的需求；此外，网络攻击事件的时效性要求很高，实时根据动态数据检测出潜在的网络安全事件，也是网络空间主动防御的难题。

专家系统可以用于提供专业的网络安全知识，并且可以根据历史网络安全事件总结出网络攻击规律，从而能有效地检测出正在发生的某些网络攻击。然而，专家系统的缺陷在于专家知识更新慢，利用专家系统能快速检测已知的网络攻击，但是对于未知的网络攻击事件，专家系统的知识往往由于更新不及时，导致系统无法正确检测。

此时，需要结合人工智能技术赋予的预测能力，对动态的数据设计在线算法，能够结合已有的网络安全知识和实时的数据判断当前的潜在网络攻击；根据已经发生的攻击事件和历史数据，建立攻击预测模型，预测未来可能发生的攻击行为，通过人工智能技术增强系统的预测能力，提供动态防御能力，提升网络安全事件的快速响应能力。

四、基于人工智能的网络空间安全防御关键技术

基于人工智能技术提升网络空间安全防御能力，需要解决从原始海量数据到有效知识的整合，人工智能技术可以通过高效的知识表示，构建网络安全知识大脑，助力实现网络安全知识综合利用和主动防御。

在已有的研究工作中，知识图谱 [15] 是一种高效的知识表示模型，虽然其在一定程度上解决了数据到知识的表示难题，但是知识图谱表示方法面临着时空特性无法有效表示、多领域知识统一表示困难等难题。MDATA 模型 [16] 通过对知识引入时间特性和空间特性，能有效解决时空特性的表示，以及支持不同领域、不同维度的安全知识的关联和融合，可用于构建大规模动态网络安全知识大脑。

基于人工智能技术构建大规模动态网络安全知识大脑，实现网络空间安全防御的关键技术主要包括网络安全知识的抽取和融合、网络安全知识表示、网络安全知识大脑构建、基于网络安全知识大脑的攻击事件研判等。

（一）网络安全知识的抽取和融合

网络安全知识的来源广泛，包括漏洞库、病毒库、告警数据、安全厂商的检测结果、安全论坛、网络安全事件报告资产描述等，为构建大规模的网络安全知识大脑，需要首先从不同来源的网络安全数据中抽取知识，并对不同领域的网络安全知识进行有效融合。

网络安全数据主要以文本数据、结构化数据、半结构化数据等类型为主，可以采用人工智能技术对数据进行抽取。例如，可以使用 word2vec 技术将文本中的单词转换为向量，结合卷积神经网络（CNN），BiLSTM，条件随机场算法（CRF）等技术进行实体和关系识别，同时按照网络安全知识的类型进行分类，并将分类以后的实体和关系添加到对应网络安全知识的实例中进行保存。

由于网络安全中用于训练的预料数据有限，可能无法覆盖所有的网络安全知识，因此需要根据已有的网络安全知识进行推理，生成新的知识。例如，已有的网络安全知识中包括某漏洞 A 的基本信息，包括受影响设备、软件、漏洞危险程度等，同时已知某资产 B 包含了对应的软件，并且未安装对应补丁，便可推理出资产 B 存在漏洞 A 的新知识。此类知识推理的方法主要包括两种，一种是自定义推理规则，根据预先制定的规则进行知识推理和演绎；

第二种是采用智能化技术，根据已有的知识进行概率推理，计算新知识存在的概率。第一种方法需要人为地定义推理规则，可扩展性较差；第二种方法使用深度神经网络进行计算新知识存在的概率，可扩展性强，但是可解释性较第一种方法差一些。

不同数据源抽取得到的网络安全知识可能会有不同的描述方式，例如很多厂商都研发了入侵检测系统（IDS），不同 IDS 系统返回的告警数据格式并不完全一致，抽取得到的入侵检测知识的描述也不一致，因此需要对网络安全知识进行有效融合。常用的融合方法包括实体对齐、基于知识表示的消歧等，基本思想是将不同的网络安全知识库按照实体和关系的统一描述进行融合。

（二）网络安全知识表示

常用的知识表示模型包括符号逻辑、语义网、专家系统、知识图谱、MDATA 模型等，通过知识表示可以将网络安全中不同类型的知识描述为统一的形式，并可通过知识的向量化进行高效计算。

知识图谱主要采用“< 实体，关系，实体 >”这种三元组形式对具体的知识进行表示，例如“<Linux kernel 漏洞，导致，DDoS 攻击 >”表示 Linux kernel 漏洞会导致 DDoS 攻击；“< Linux kernel 5.1.13，存在，Linux kernel 漏洞 >”表示 Linux kernel 5.1.13 的版本存在该漏洞。知识图谱能有效描述网络安全知识，但是当知识动态变化时，对应的三元组及相关联的知识很难及时更新。

MDATA 模型对实体之间的关系、属性的时空特性进行表达，从而有效表示网络安全知识的动态变化情况。具体而言，在关系和实体属性上增加了时间和空间特性，如某系统存在漏洞的知识，添加存在漏洞的时间区间，从而能更详细地表示系统的实际安全情况。网络攻击可能通过不同的 IP（网络之间互连的协议）地址等，攻击事件中的 IP 地址等特性则作为网络安全知识中的空间特性。MDATA 模型通过对时间、空间特性的描述，可以表示出网络安全知识的动态变化过程。

（三）网络安全知识大脑构建

网络安全知识大脑的构建包括两部分：网络安全知识库（SeKG）和场景知识库（ScKG）。其中，网络安全知识库是通用的网络安全知识的集合，并且可以随时或定期更新补充；而场景知识库是特定知识的集合，可以依据仿真攻击的设定而定，也是描述具体攻击行为的知识库。

网络安全知识库和场景知识库可以根据概念、实例、关系、属性、规则的五元组模型进行构建 [18]。其中，概念是抽象本体的集合，如操作系统、软件、攻击等；实例是具体例子的集合，如 Windows 7，Adobe Reader，DDoS 等；关系表示实例之间存在的关系，如 subClassOf，instanceOf，is a (ISA) 等；属性包括实例属性值的集合；规则用来推演新的属性值和新的关系。

构建网络安全知识大脑用到的概念主要有漏洞、资产、软件、操作系统和攻击等。其中，漏洞信息来源于漏洞库，每一个漏洞都有唯一的身份标识号（ID）和类别标识。资产则包括软件和操作系统等，软件和操作系统主要涵盖当前市面上使用的所有版本。攻击主要是针对利用漏洞的攻击，攻击的信息也主要是来源于漏洞库，因为漏洞库里对漏洞的描述会包含很多详细的信息，包括漏洞会导致哪些攻击发生等。

（四）基于网络安全知识大脑的攻击事件研判

网络空间防御面临的主要威胁是网络攻击，一般而言网络攻击可以分为单步攻击和复合攻击。单步攻击可以理解为针对某资产发动的离散的攻击，而复合攻击可以理解为是有多个单步攻击排列组合而成的，也就是说复合攻击有多个攻击步骤，而这些攻击步骤之间是有关系的，不是离散的、无关联的，攻击步骤之间有因果关系、顺承关系、选择关系等。

单步攻击的研判相对而言简单，已有的基于规则、特征的检测方法能取得很高的成功率。而复合攻击的检测难度大，典型的复合攻击包括 APT 攻击等。复合攻击通常是以攻击链的形式发生的，可以看作是多个单步攻击的排列组合。不同的操作系统上会安装不同的应用软件，不同的应用软件会有不同的漏洞，也会感染不同的木马，而这些木马和漏洞会导致相同或不同的单步攻击，此外，一些操作行为（网络、注册表、进程和文件）也会导致相同或不同的单步攻击，入侵检测系统会产生安全事件，这些安全事件就是一个单步攻击，而每一个单步攻击都属于攻击链中的某一类，所有的单步攻击

根据产生的效果进行排列组合就形成了不同的复合攻击，而排列组合的和攻击链中的时序、依赖关系等高度相关。

使用网络安全知识大脑研判网络攻击时，可利用有限状态机 [19]，设置初始状态、中间状态、终止状态和触发条件，并添加容错机制，可以在缺失数据的时候仍然生成复合攻击的攻击链，在网络安全知识库和场景知识库的基础上，描述复合攻击的各个步骤之间的关系，然后根据攻击步骤的关系、时间先后关系、IP 的传播关系等来判断是否可以生成攻击链。如果满足，则输出复合攻击的攻击链，如果不满足，就去知识库中查找等价的步骤，或补充生成攻击链并输出。当输入的数据中存在误报和漏报的情况时，基于网络安全知识大脑的研判可以自动补全缺失的信息，生成一条完整的攻击链，从而提高攻击研判的准确率，为网络安全主动防御提供支撑。

五、发展对策与建议

随着人工智能时代的到来，网络空间安全面临着很多新风险和新挑战。将人工智能技术应用在网络安全防御中，可以大幅度提升网络空间防御能力。具体发展对策与建议如下。

（一）构建动态可扩展的网络安全知识大脑

充分利用人工智能技术在处理海量数据、多源异构数据、实时动态数据等方面的显著优势，构建动态可扩展的网络安全知识大脑，提升网络空间防御能力。具体而言，针对网络安全知识描述中多实体、弱关系、时空复杂性和多来源等特点，对于结构化、半结构化和非结构化的数据，基于 MDATA 知识表示模型、网络安全知识语料库中的特定表达和网络安全知识之间特有的逻辑关系和对应关系，构建相应的网络安全本体模型，实现多领域知识的统一表示，相较于传统的知识图谱等知识表示模型，可提升融合效率和多领域动态知识统一表示的准确率。

在此基础上，针对半结构化数据和非结构化数据知识抽取难的问题，基于构建的本体模型，结合双向循环神经网络和条件随机场等深度学习方法，进行特征抽取、联合标记、类别标记等。对于未被识别出的本体进行人工抽取，从而确保基于本体模型生成的三元组知识在逻辑上是正确的，实现动态可扩展的网络安全知识大脑，为网络安全防御提供强大的具有自学习能力的知识库支撑。

（二）推动有效网络攻击的智能化检测

针对网络攻击越来越智能化，大规模网络攻击越来越频繁，网络攻击的隐蔽性越来越高的特点，遵循网络攻击的基本规律，基于构建的网络安全知识库，实现复杂网络攻击的智能化检测算法。可将网络安全事件和攻击的基本信息存储于安全知识图谱中，通过采集数据与安全知识图谱进行匹配，根据状态的触发约束，可分析得到网络系统中的单步攻击和复合攻击。

一般而言，检测复合攻击时可以分析出攻击的当前阶段，无法确保输出完整攻击链，因此可进一步基于攻击规则库的复合攻击研判技术，将先验知识存入网络安全知识图谱和复合攻击规则库，基于大数据分析平台，通过采集数据与安全知识图谱进行匹配，再经过时空属性和复合攻击规则库的共同约束，从海量的数据中挖掘出有效的攻击链，并完善复合攻击的攻击链，实现自动化分析攻击目的和意图等。针对传统方法无法应对输入的数据中误报和漏报的情况，可以基于多模态数据的复合攻击研判，当输入的数据中存在误报和漏报的情况时，可以自动补全缺失的信息，并计算生成不同攻击链的概率，消除误报和漏报的影响。进一步，可通过网络仿真平台对攻击事件进行仿真，将分析结果与仿真攻击的信息进行对比，实现对有效网络攻击的智能研判。

（三）评估人工智能技术的安全性，推动人工智能技术的良性应用

目前人工智能技术还不能完全脱离人而存在，人的引导至关重要。在制定人工智能的发展路线的同时应该要紧盯风险防御，加强对潜在风险的预判和研究，注重系统安全防御技术的发展，明确防御发展策略。不能盲目地将人工智能作为一项“百利而无一害”的技术进行使用，在进行顶层设计的同时考虑风险管理，对人工智能技术的安全性进行有效评估，为人工智能乃至系统防御技术提供有效规范的引领作用。

同时，应加强人工智能风险管理。人工智能自身存在的漏洞和人工智能技术的滥用是系统安全防御中很难避免的环节。自身存在的安全风险属于最致命的问题，应用越广泛，其带来的危害性也越大。系统安全防御技术要从人工智能技术自身入手，构建主动免疫的计算构架，尽可能地降低技术自身的漏洞危害，不断创新保持技术优势。

六、结语

人工智能既能用来提升网络空间安全，又会带来新的风险与挑战。基于人工智能技术提升网络空间主动防御能力，是保障网络空间安全的重要途径。为此，需加强人工智能用于网络空间安全防御关键技术的研究，构建大规模动态网络安全知识大脑，推动有效网络攻击的智能化检测，加快评估人工智能技术的安全性，推动人工智能技术在网络空间领域的良性发展与应用，全面提升我国网络空间安全保障能力。

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 security and safety [M]. Beijing: Publishing House of Electronics Industry, 2020.

[2] 贾焰, 方滨兴. 网络安全态势感知 [M]. 北京：电子工业出版社，2020.
Jia Y, Fang B X. Network security situation awareness [M]. Beijing: Publishing House of Electronics Industry, 2020.

[3] Veeramachaneni K, Arnaldo I, Korrapati V, et al. AI^2: Training a big data machine to defend [C]. New York: IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS), 2016.

[4] Hendler D, Kels S, Rubin A. Detecting malicious powershell commands using deep neural networks [C]. Incheon: ACM Asia Conference on Computer and Communications Security, 2018.

[5] Manès V J M, Han H, Han C, et al. The art, science, and engineering of fuzzing: A survey [J]. IEEE Transactions on Software Engineering, 2019. doi: 10.1109/TSE.2019.2946563.

[6] Brooks T N. Survey of automated vulnerability detection and exploit generation techniques in cyber reasoning systems [EB/OL]. (2017-02-20) [2021-02-28]. https://arxiv.org/abs/1702.06162.

[7] Capture the Flag [EB/OL]. [2021-02-28]. https://defcon.org/html/defcon-24/dc-24-ctf.html (2016 Defcon CTF Final Scores).

[8] Fortinet FortiGuard Labs 2018 Threat Landscape Predictions [EB/OL]. (2017-11-14) [2021-02-28]. https://www.fortinet.com/blog/business-and-technology/fortinet-fortiguard-2018-threat-landscape-predictions.html (Prediction: The rise of Hivenets and Swarmbots).

[9] Kirat D, Jang J Y, Stoecklin M P. DeepLocker-concealing targeted attacks with AI locksmithing [EB/OL]. (2018-08-09) [2021-02-28]. https://i.blackhat.com/us-18/Thu-August-9/us-18-Kirat-DeepLocker-Concealing-Targeted-Attacks-with-AI-Locksmithing.pdf.

[10] Hu W W, Tan Y. Generating adversarial malware examples for black-box attacks based on gAN [DB/OL]. (2017-02-20) [2021-02-28]. https://arxiv.org/pdf/1702.05983.pdf.

[11] 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 (Early Access), 2020. doi: 10.1109/TNSE.2020.2996738.

[12] Gu Z Q, Cai Y Y, Wang S, et al. Adversarial Attacks on Content-Based Filtering Journal Recommender Systems [J]. Computers, Materials & Continua, 2020, 64(3): 1755–1770.

[13] Shokri R, Stronati M, Song C Z, et al. Membership inference attacks against machine learning models [C]. San Jose: IEEE Symposium on Security and Privacy, 2017.

[14] Zhang Y H, Jia R X, Pei H Z, et al. The secret revealer: Generative model-inversion attacks against deep neural networks [C]. Seattle: IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[15] Ji S, Pan S, Cambria E, et al. A survey on knowledge graphs: Representation, acquisition and applications. [DB/OL]. (2020-02-02)[2021-02-28]. https://arxiv.org/abs/2002.00388.

[16] Jia Y, Gu Z Q, Li A, et al. (Eds). MDATA: A new knowledge representation model [M]. Switzerland: Springer International Publishing, 2021.

[17] Hinton G E. Learning multiple layers of representation [J]. Trends in Cognitive Sciences, 2007, 11(10): 428–434.

[18] Jia Y, Qi Y, Shang H, et al. A practical approach to constructing a knowledge graph for cybersecurity [J]. Engineering, 2018, 4(1):53–60.

[19] Qi Y, Zhong J, Jiang R, et al. FSM-based cyber security status analysis method [C]. Hangzhou: IEEE Fourth International Conference on Data Science in Cyberspace (DSC), 2019.