# 人工智能在网络安全领域的应用现状

刘文懋
绿盟科技

关键词：人工智能　网络安全

人工智能并不是近年来的新概念，自上世纪50年代起就已有人工智能的相关研究了。随着相关技术的不断突破，人工智能在数十年的发展历程中也出现了数次高峰波谷，而近年来深度学习应用大获成功，开始推动人工智能在很多行业的前进。当前在某些领域，如图像识别、棋类竞技，人工智能已经演进到第三代，有了超越大部分人类的智能水平，甚至学术界已经开始讨论“强人工智能”，也就是能自我推理和决策的智能了。

人工智能是否能应用在网络安全领域？这是一个非常值得探讨的问题。事实上，网络安全的本质在于攻防双方之间的对抗，而棋类竞技本质也是棋手之间的博弈，两者在某些方面存在共通之处。众所周知，以AlphaGo/AlphaGo Zero为代表的对抗学习技术，已经能成功挑战人类顶尖棋手。此外，三星、脸书（Facebook）以及中科院自动化所分别以95.91%、90.86%、87.11%的胜率在2018年“星际争霸AI挑战赛”中荣获前三名[1]。将人工智能技术应用在博弈对抗的领域似乎非常有前景，人工智能在网络安全领域的成功应用似乎也指日可待。

## 人工智能在网络安全领域应用的挑战

过去几年，研究者试图将人工智能应用在网络安全领域，以解决若干问题。但从实践过程和结果来看，还存在巨大的挑战。接下来，我们从检测、溯源、认知、决策等方面，依次进行分析。

### 攻击者绕开检测特征，产生漏报

世界上的主要科技强国，包括中美两国，都将网络空间安全纳入到国家安全的范畴，也成立了相应的网络安全部队，换言之，网络空间对抗的最高形态，已无异于战争。孙子曰：“兵者，诡道也。”军事虽有理论支撑，但兵法运用之妙，存乎一心。真实战争不存在定式，无论是物理形态的战争还是网络空间对抗，攻击者不会遵从对方的防守体系，或者按照防守方的思路去层层突破，从古到今以弱胜强的经典战役均是出其不意、攻其不备，找到对方的弱点和漏洞，重点突破。

更何况，如今国家支持的威胁（state-sponsored threat），已经超越了地理或物理的限制，攻击方将未知漏洞纳入武器库，持续潜伏，伺机一击必杀。越是对抗高的场景，检测引擎越容易被攻击者绕过。本质上人工智能将模型特征替换了规则，但如果攻击者的恶意行为模式在当前的人工智能算法选择的特征集之外，就有可能绕过这些算法引擎，形成“降维打击”。

举一个简单的例子。企业中普遍使用网络侧的安全检测和防护机制，但现在攻击者通常会使用加密技术使恶意软件与主控端（C&C）通信以实现持久

化，因而即便使用网络侧人工智能能够识别一些规则无法覆盖的新型攻击载荷，但对于加密流量则难以生效。又如，为了躲避各类网络和终端的安全探针，在近年的各类大型攻防演练中，攻击者倾向于采用前期钓鱼、社会工程（库）等方式获得内部员工的合法身份，进而在业务层窃取数据或横向移动，导致在后期，所有网络层面或终端层面的人工智能检测引擎无能为力（因为没有网络或操作系统层面的恶意攻击行为）。事实上，每年攻防演练的情势都不同，被动地补齐上一年场景中的检测能力，效果不会尽如人意。

正所谓“道高一尺，魔高一丈”，攻防永远是技术、思路的对抗博弈，期望人工智能在某个细分领域获得成功以解决所有问题的思路是不切实际的，这也是当前体系化安全大脑尚不成熟的重要原因。

## 概念漂移，多场景检测率低

深度学习在工业界的很多应用（例如图像识别）中性能优异，得益于海量样本的训练。在学术界，从事人工智能的研究者通常可以根据某个特定场景，设计一种有针对性的算法和模型，然后针对某个公开数据集或私有场景获得的数据集调整模型，以获得良好的性能。

然而，对于网络安全中的样本学习，最大的挑战在于缺乏标记的样本，因为缺乏有经验的安全人员，内部环境中的攻击事件也很少。我们可以针对某次对抗演练，人工地将探针数据划分为训练集和测试集，然后在这个数据基础上训练得到模型参数，最终验证得到很好的检测准确率和召回率。但是，该场景黑白样本的绝对数量还是太少，原因是当前安全专家太少，无法对网络、终端和应用行为进行大规模标记，而一般水平的安全运营者缺乏标记能力，这与人类具有普遍认知能力的图像识别场景截然不同。这种情况导致该场景的模型参数可能在其他演练场景下性能非常糟糕，也就是概念漂移[2]，其原因也很直观：

1. 攻击者会时常调整攻击手法，即便方法类似，具体攻击载荷可能与前一次存在很大差异，现有模型可能会有漏报。

2. 不同机构的业务差异很大，训练环境中的白样本与测试环境的白样本不同，导致黑白样本的分界线产生偏移。

## 溯源图依赖爆炸，还原攻击路径困难

在网络空间战场中，攻击者的行为是复杂多变的。在确定攻击事件后溯源攻击者的攻击路径，对安全运营人员来说是十分必要的。溯源如同大海捞针，困难重重，其中最大的挑战在于溯源图过于庞大，难以找到攻击者关键的攻击路径。

笔者团队在一个靶场环境中，先通过文件漏洞将蚁剑 Webshell[1] 上传到服务器；然后，利用 Webshell 连接靶机虚拟终端采集信息并提权（提高自己在服务器中的权限以便控制全局）；接着，实现对靶机的持久化控制，并以该靶机为出发点进行内网横向移动，如图 1 所示。针对这种攻击模式，结合网络侧与终端侧数据构建有效的溯源图是进行攻击溯源的关键。溯源图主要是挖掘进程、文件、IP、注册表、服务等实体之间的依赖关系。这种依赖关系在正常用户行为中也存在。与正常用户行为相比，攻击者的攻击路径只占整个溯源图的极小部分。以前述场景为例，溯源图包含了 1000 多个顶点与 200 多万条边，而安全运营人员关注的仅仅是图 1 中简单的攻击路径。因此，攻击溯源首先要解决的问题就是从复杂的大规模溯源图中找到攻击者的攻击路径，也就是通常所说的依赖爆炸问题，这给溯源带来了很大的挑战。

## 如何认知，何为知识

笔者团队曾经通过无监督学习建立业务基线、分析攻击手法相似度等方法，检测到了若干告警，通过告警又溯源出疑似攻击路径，但这些路径是否为真实的攻击路径，经过专家判断后，发现最终验证效果并不理想。

[1] Webshell 是一种攻击技术，攻击者可以上传 Webshell 恶意网页，如蚁剑 Webshell 文件内容为“<?php @eval($_POST[‘key’]);?>”，然后访问该网页并在页面执行命令，其作用与系统 Shell 执行命令一致。

原因在于这些算法只有“智能”，不体现“人工”，很多疑似告警或告警路径只存在相关性，而非真正需要处置的告警。后来我们引入了攻防专家，通过识别攻击意图，例如区分探测性还是利用性，最终推荐出 TOP 10 告警（正好是安全团队半天的处理量），取得了很好的效果，因为这些告警确实是安全团队需要及时处理的。

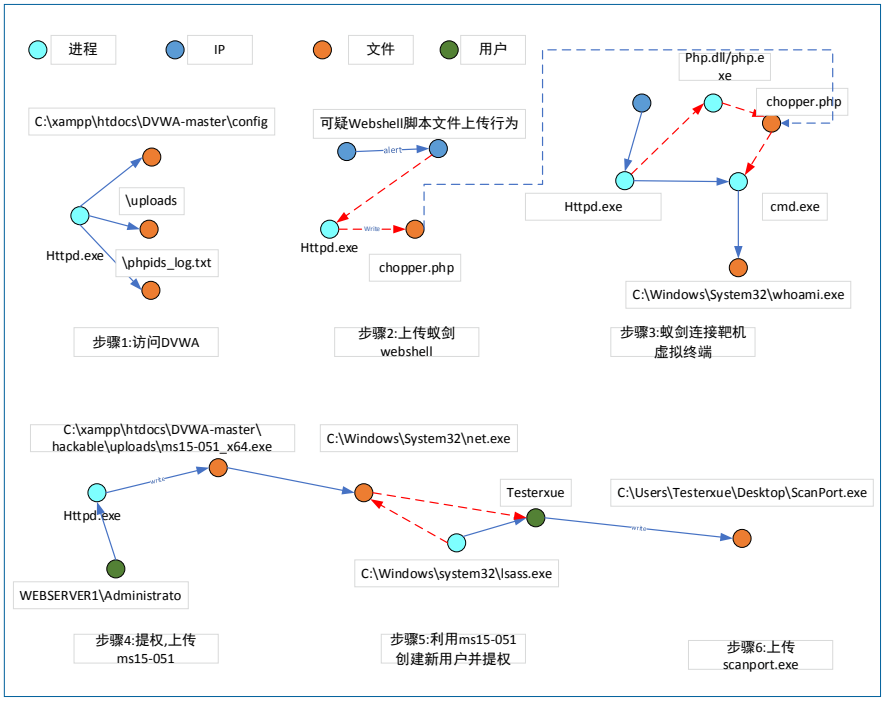


图 1　Webshell 文件上传及内网横向移动场景图

当前，我们能够通过人工智能的算法识别出一些关键告警，但要达到自我认知的水平，则需要生成标准的 TTP[2]（Tactics, Techniques and Procedures，战术、技术和过程）的威胁情报元数据（Indicator of Comprise，IoC），而高层级的情报（例如攻击链和攻击团伙）则需要依次关联告警、事件、攻击路径，最后到攻击者，每层关联都需要引入额外的知识，例如 MITRE（一个向美国政府提供系统工程、研究开发和信息技术支持的非营利性组织）的 ATT&CK[3]、CAPEC[4] 和 CVE[5] 都是潜在的知识库。

无论是攻击意图识别，还是 MITRE 的知识库，当前阶段还都不能直接帮助人工智能进行认知，其中还存在巨大的鸿沟。当前这些知识库的主要用途是解释已经发生的攻击行为，例如还原 APT（高级可持续威胁攻击）事件的攻击链，但不能根据知识库自主地推导未知的攻击链。

## 智能决策困境，知其然不知其所以然

对于安全运营团队而言，理想的人工智能是能够打通认知、溯源、预测和决策多个环节，形成端到端的安全事件智能处置。从结果看，即便人工智能算法在认知、溯源和预测环节犯错，也只是增加漏报和误报，即增加安全团队的工作量。但在决策环节，人工智能犯错则会造成难以挽回的后果，例如 WAF（Web 应用防火墙）错误的策略会导致网站业务中断，防火墙错误的规则会导致断网，如果在工控、车联网场景下，其后果更为严重。

因此，可靠性（reliability）、可用性（availability）和安全性（safety）是人工智能决策算法压倒性的指标，要实现可用的智能决策引擎，则需要满足两个要点：

1. 需要有可证实、可解释的证据和推导链，而不能只是通过深度神经网络产生的“感觉”。无法解释的决策是不可用的，更是不值得信任的。

2. 需要了解 IT 系统中的资产重要性和处置产生的结果。不同的环境，即便遭遇同样的威胁，其决策

[2] TTP 通常用于解释攻击者的具体攻击模式，战术（Tactics）有前述的 Webshell 或口令爆破等，技术（Techniques）有网络加密等，过程（Procedures）有攻击链中的侦查阶段等。

结果也是不同的。

当前，即便让人工智能算法通过采用更深的网络、嵌入更多特征、加入更多样本，能达到 90% 甚至更高的准确率，但说不清依据，就无法应用于自主决策，这是智能决策的最大困境。

## 人工智能：能否超越人工的智能

当五年前深度学习开始兴起时，网络安全从业者就在畅想：基于大数据、人工智能的新型检测技术是否能超越传统的入侵检测技术，发现过去很难发现的网络威胁呢?

网络安全产业面临的一个很大的挑战是缺乏专业的人才，当前国内网络安全人才缺口近百万，我国每年网络安全人才的培养数量远远不足以弥补这个缺口[6]。当前将人工智能应用于网络安全最大的驱动力是通过算法学习将顶级专家的知识融入模型，从而通过规模化部署降低整体专家的边际成本。因而，在这种思路下，人工智能得以最大程度地接近“人工”（也就是专家）的智能水平。

至于未来，人工智能能否帮助我们发现未知的漏洞或未知的威胁犹未可知，例如通过智能的模糊测试、自动攻击系统找到深层次的未知漏洞，进行自发的多步攻击；或者通过学习知识库，自主推导攻击行为之间的关联，从海量数据中发现新型的攻击，甚至能自主进行防护等。10 年或 20 年后，人工智能必然是要超越顶尖人类攻防专家的，事实上攻防武器化已成为一个重点发展趋势，但在此前，需要经过持续的知识积累。对于攻方，需要补充资产、漏洞和利用（exploit）之间的知识；对于守方，需要在现有知识库的基础上，增加可关联性、准确语义等要素。

## 小结

本质来看，当前阶段人工智能在网络安全领域的应用还充满挑战，主要是因为人工智能缺少网络对抗的相关知识，有的知识可以通过日常运营进行丰富并最终接近人类专家的水平，而有的知识则超越了当前防守方已有的知识体系，而强人工智能显然还无法实现真正的全网络空间知识“自学习”。

# 人工智能赋能网络安全的典型应用

尽管人工智能在网络安全领域全面应用存在诸多挑战，但我们还是在一些应用中看到了成功的曙光。

## 特定领域的异常检测

在网络安全领域，人工智能比较适合解决一个特定的问题，例如某种异常行为的检测。如果融入了专家知识，那么与其他的异常检测问题没有区别。下面将介绍笔者团队正在从事的几项基于机器学习的异常检测机制，包括 Webshell 异常检测、加密流量识别、DGA 恶意域名检测[7] 等。

## Webshell 异常检测

Webshell 是一种集成了探测、利用、持久化和进一步攻击的常见攻击手段，特别是攻陷 Web 服务器后执行的指令，应该立刻得到安全团队的关注。但由于以往 WAF 和 IDPS（入侵检测防御系统）是规则驱动的，往往会产生漏报。因而现有一些工作[8] 利用 Webshell 流量数据进行特征挖掘分析，构建流量数据的特征向量，采用监督学习算法对异常流量进行检测，从而对 Webshell 进行分类识别。

不过 Webshell 本身具有多种意图，攻击者在多个攻击阶段使用的是同一大类的攻击载荷，因而以往区分攻陷事件的手段往往会引起误报。因此，我们在构建特征覆盖面的前提下保证检测的精确率和召回率。在特征工程阶段，结合 Webshell 的特点和相关的专家知识去挖掘信息，例如存在系统调用的命令执行、文件操作函数（如 eval、system、fopen 等），以及伪装性很强的加解密函数等衍变方法（如 postbody 长度、KV 数量，特殊字符长度、关键字数目等）。

在现有的一些研究实验中，对 Webshell 流量的识别准确率往往比较高，能够达到 90% 以上。

最终实验的训练样本约 104 万，共分为 10 类，测试样本约 26 万，相关的测试结果如表 1 所示。目前单模型整体效果最好的是 MLP；ML（见表中的 GBDT 和 RF）和 DNN（见表中的 LSTM 和 TextCNN）模型表现相当。

表 1　基于机器学习的 Webshell 检测实验结果

| 名称 | LSTM | MLP | TextCNN | GBDT | RF |
|---|---|---|---|---|---|
| 精确率 | 0.98630 | 0.99911 | 0.99903 | 0.99682 | 0.99250 |
| 召回率 | 0.98459 | 0.99972 | 0.99903 | 0.99966 | 0.95456 |
| F1 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |

## 加密流量识别

高德纳咨询公司（Gartner）曾经预计，在 2019 年，多于 80% 的 Web 流量会被加密，到 2020 年，超过 60% 的组织解密 HTTPS 流量会失败，以致错失定向的 Web 恶意软件。通常企业将自己的私钥交给安全企业或安全设备以解密加密流量是非常困难的，因而加密流量的识别将成为一个非常重要的应对机制。

除了常见的加密代理识别外，恶意流量和恶意软件的识别也是重要的研究方向。还是以 Webshell 为例，当前很多工具（如冰蝎、哥斯拉等）为了躲避 WAF 检测，开始使用加密技术作为命令通信通道。我们提取了数据包载荷的一些头部特征，以及数据流的一些统计特征，使用 LightGBM（一种梯度提升模型）为分类模型，验证了该模型对冰蝎检测有效。这些载荷特征和统计特征可以准确刻画冰蝎的特征，无须调整模型就可以检测新的冰蝎版本（3.0）[9]。

更进一步，该模型融合上节提及的特征，可以检测非加密的 Webshell、非加密的 Webshell over HTTPS，以及加密的 Webshell 三类恶意流量。

此外，在针对恶意软件通过加密通道传输的数据进行检测方面，业界同样有一些积极结果，因篇幅所限不做详述，读者可参阅文献 [9]。

## AI 辅助安全运营（AISecOps）

在很长一段时间内，以设备为中心的安全运营基本能满足企业要求，但随着 APT 威胁持续变强和对抗演练成为常态，安全团队已经无法仅仅依靠安全设备的告警保障网络空间安全，一方面各类安全设备功能各异，互为补充，需要将各类告警进行聚合才能还原整个攻击链；另一方面，安全设备产出大量低危、试探性质的告警或误报，安全团队无法在短时间内进行有效处理。

AI 辅助安全运营（AISecOps）[10] 试图解决这个问题：首先，依靠人工智能的学习能力，识别关键告警，进而将相关告警进行关联；其次，对历史告警进行溯源，对未来告警进行预测；最终，根据当前态势制定行动剧本（playbook），通过统一的安全控制器下发策略进行安全编排（orchestration）。

借鉴自动驾驶的分级模型，我们对 AISecOps 的成熟度进行了分级，如表 2 所示，其中颜色越深代表越成熟。然而，人工智能还不能超越顶级专家，网络安全在某些关键场景中的处置至关重要，所以幻想在所有场景下能实现 L5 级的完全人工智能化运营是不切实际的，但借助 AI 技术大幅降低专家的边际成本是完全可能的。

# 网络安全领域人工智能的发展趋势

经过数年实践，笔者认为在网络安全领域，人工智能的未来发展趋势如下。

## 可解释人工智能

如果将人工智能用于图像处理，并不需要太多推理解释，因为图像中是什么内容，人类一眼就能看懂，即便图像有一些像素损失也不影响最终的识别结果，因为图像具有天然的可解释性。然而，网络安全领域通常是攻击载荷、规则和对应的触发条件，请求中的字段有一个字节的差异，最终的结论也会大相径庭。深度学习技术的“知其然不知其所以然”的弱点，在重证据、重报告的对抗场景下，不足以说服安全运营团队采用。

正因如此，可解释人工智能（eXplainable AI，XAI）近年来成为学术界一个新的研究方向，在网络安全领域显得尤为重要。

以前述 Webshell 异常检测为例，我们使用 LIME 内核解释训练所得模型，以样本“`alert tcp any any -> any 80 (sid:900001; content:"z1";content:"base64 _ decode";http _`

client _ body;flow:to _ server,established; content:"**POST**"; nocase;http _ method; ;msg:"Webshell Detected Apache";)" 为例，可以得到图 2 的解释结果，可见 z1、eval 等关键词是其为恶意的主要原因，进而我们可以将其作为知识，形成确定的、令人信服的规则。

更通用地，笔者团队研究了一种基于可解释人工智能的序列分析算法 [11]，自动化地提取模型学习到的特征规则，有效降低恶意样本特征规则提取对专家的依赖。目前主要是针对明文流量或者文本之类的恶意样本,进行特征提取的自动化。目前该工作已开源 [12]，感兴趣的读者可参考，欢迎一起参与扩展场景。

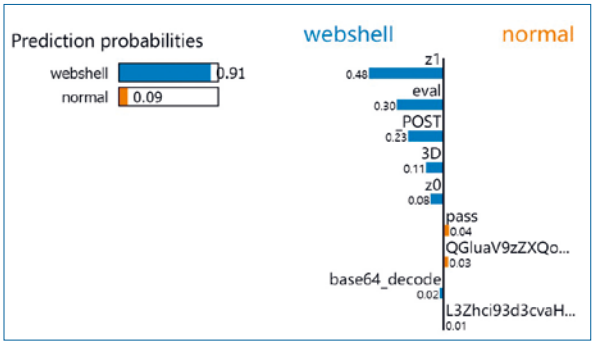


图 2　Webshell 异常检测解释结果

## 攻防专家 + 人工智能融合

笔者所在的 AISecOps 团队在四年前就开始使用机器学习做异常检测，例如用户行为分析（User &

表 2　AISecOps 的成熟度分级

| 自动化水平 | 名称 | 定义 | 任务阶段 | | | | | | | | | 数据交互（DIKW 模型） |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 感知阶段 | | 认知阶段 | | | 决策阶段 | | 行动阶段 | | |
| | | | 识别 | 检测 | 关联 | 溯源 | 预测 | 评估 | 制定 | 响应 | 反馈 | |
| L0 | 无自动化 | 由运营人员全权完成安全运营操作 | | | | | | | | | | 数据采集 |
| L1 | 运营辅助 | 自动化运营系统完成感知、认知、决策中的多个子任务，其他运营操作由人完成 | | | | | | | | | | 数据集成<br>信息加工 |
| L2 | 部分自动化 | 自动化运营系统针对指定初级任务完成感知、认知、决策、行动全流程子任务，与运营人员进行持续数据交互 | | | | | | | | | | 信息融合<br>知识获取 |
| L3 | 有条件自动化 | 自动化运营系统完成包含行动层子任务在内的全流程子任务，运营人员须在关键阶段提供适当应答 | | | | | | | | | | 知识理解<br>知识沉淀 |
| L4 | 高度自动化 | 在限定场景下，自动化运营系统完成包含行动层子任务在内的全流程子任务，运营人员不一定提供应答 | | | | | | | | | | |
| L5 | 完全自动化 | 在所有场景下，自动化运营系统完成包含行动层子任务在内的全流程子任务，运营人员不一定提供应答 | | | | | | | | | | |

Entity Behavior Analytics, UEBA），在实验环境中能检测出预期的异常场景，但在实际场景中，却获得了大量未预期的告警，大部分是误报。总结经验教训，数据驱动安全虽是新方向，但没有攻防内核的人工智能丢失了安全的本质：对抗。

因而，我们团队加入了具有多年攻防经验的成员，认真总结攻击者的手法，例如 Webshell 的试探性和利用性告警的差别，加入关键告警的语义特征，从而得到了比较满意的结果。

可以预见，未来安全运营的正确步骤是：首先总结安全团队的经验，然后通过人工智能技术将经验融入到模型，最后形成特定场景下可解释的检测引擎。而不是相反的步骤。

## 可推理的知识生成

在攻击溯源和智能决策中，需要利用感知层检测到的关键告警，但告警如何处理才能形成溯源证据链或最终决策，还是很大的空白。从逻辑上看，告警的处理依赖于一个庞大的知识库，但是当前的各类知识库还远远不能满足要求，因为它们都还只是用于表示，不能用于推理。如果不能推理，就只能让专家进行人工研判，无法发挥人工智能的优势进行自主学习和推理，也无法降低运营的成本。

如果要达到 L4/L5 级的自动化安全运营目标，就应该研究如何构建可推理的知识库，包括知识库的语法、语义模式，以及知识库的推理机制。理想状态下，人工智能能够自主地学习当前知识库条目，关联、创造不同的知识库条目；能够根据新的漏洞、威胁情报、资产自主地在知识库中添加新的条目；在运行时能够根据检测到的告警构建知识子图，包括当前、历史和未来受影响的资产、攻击手法以及所利用的脆弱性，以自然语言的形式形成最终报告和处置建议。

## 结论

总体而言，人工智能的成功一定能在网络安全领域得到复制，但前提是需要解决其中的若干挑战。人工智能当前可以在一些网络安全领域的特定问题中得到较好的应用，后续自动化地辅助安全运营是未来成功的关键。 ■

**致谢：**本文中的图表、数据均来自绿盟科技天枢实验室和创新中心的研究成果。

**刘文懋**

CCF 高级会员、理事。绿盟科技创新中心总监，星云实验室负责人。主要研究方向为云计算安全、网络安全。
liuwenmao@nsfocus.com

## 参考文献

[1] 中科院自动化所智能系统与工程研究中心. 2018 年星际争霸 AI 挑战赛中科院自动化所夺得季军，三星与 FB 获冠亚军 [OL].http://www.crise.ia.ac.cn/news_view.aspx?TypeId=28&Id=437&FId=t2:28:2.

[2] 绿盟科技天枢实验室. 模型又不适用了？——论安全应用的概念漂移样本检测[OL].https://mp.weixin.qq.com/s/XFuv0orQ61XhrI2O-CyJSw.

[3] MITRE. ATT&CK[OL].https://attack.mitre.org/.

[4] MITRE.CAPEC[OL].https://capec.mitre.org/.

[5] MITRE.CVE[OL].https://cve.mitre.org/.

[6] 中国教育网络 .ISC2018 互联网安全大会热议网络安全人才培养 [OL].(2018-11-23). https://www.sohu.com/a/277449464_278960.

[7] Tong M, Zhang R. Far from classification algorithm: dive into the preprocessing stage in DGA detection[C]// *Proceedings of the IEEE Trustcom 2020*.

[8] 胡必伟. 基于决策树的 Webshell 检测方法研究 [J]. 网络与通信, 2016(6).

[9] 王萌. 关于恶意软件加密流量检测的思考 [OL].https://mp.weixin.qq.com/s?__biz=MzIyODYzNTU2OA==&mid=2247489152&idx=1&sn=fbe9a42e889e78c19e593d2dbbdbbe35.

[10]绿盟科技. AISecOps 智能安全运营技术白皮书 [OL]. https://www.nsfocus.com.cn/html/2020/92_1218/142.html , 2020.

[11]Zhang R, Tong M, Chen L, et al. CMIRGen: Automatic Signature Generation Algorithm for Malicious Network Traffic[C]// *Proceedings of the IEEE Trustcom 2020*.

[12] https://github.com/oasiszrz/XAIGen