前　言

霍　玮[1,4], 梁　彬[2], 张　磊[3], 葛仕明[1,4]

[1]中国科学院信息工程研究所 北京 中国 100093
[2]中国人民大学 北京 中国 100872
[3]中国科学院计算技术研究所 北京 中国 100190
[4]中国科学院大学 网络空间安全学院 北京 中国 100049

随着以深度学习为代表的新一代人工智能技术的涌现，网络空间安全局势正面临高速变化带来的各类机遇和挑战，也带来更多问题和风险。人工智能将成为网络空间安全的全新战场，如何保障人工智能算法和应用的安全性、以及如何应用人工智能技术提升系统安全水平，已成为当前国内外科研和产业关注的重点。本期专题旨在总结当前国内外研究趋势, 并展示国内研究人员通过智能化手段提升系统安全与人工智能本身安全性分析方向的最新研究成果。

本期专题从众多稿件中遴选出 9 篇收录。每篇稿件均经过多次审稿与复审。专题收录 3 篇人工智能系统安全分析的文章，它们分别从模型可解释性、模型鲁棒性以及专用处理器安全性的角度探讨了人工智能系统存在的潜在安全风险；并收录 3 篇利用机器学习技术提高系统安全性的文章，它们分别将机器学习技术应用到恶意域名识别、隐藏信息识别以及跳频信号识别等方面；最后收录了 3 篇系统安全相关文章，它们针对当前领域热点智能合约以及版权保护分别提出了新颖的漏洞检测技术和信息隐藏技术。

《深度学习模型可解释性的研究进展》对深度学习模型可解释性的研究进展进行系统性的调研，从可解释性原理的角度对现有方法进行分类，并且结合可解释性方法在人工智能领域的实际应用，分析目前可解释性研究存在的问题，以及深度学习模型可解释性的发展趋势。为全面掌握模型可解释性的研究进展以及未来的研究方向提供新的思路。

《人工智能对抗环境下的模型鲁棒性研究综述》总结了现阶段在人工智能对抗环境下的模型鲁棒性研究，论述了当前主流的模型鲁棒性的研究方法，从一个比较全面的视角探讨了对抗环境下的模型鲁棒性这一研究方向的进展，并且提出了一些未来的研究方向。

《一种针对多核神经网络处理器的窃取攻击》提出了一种针对多核 CNN 处理器的用户算法信息窃取攻击方法，并针对多核神经网络处理器在时间和内存侧信道方面的脆弱性，提出了有效的防御手段，对如何保护神经网络处理器的安全提供了一定的参考意义。

《基于机器学习的僵尸网络 DGA 域名检测系统设计与实现》设计实现了一种 DGA 域名检测系统。基于随机森林算法的轻量级分类分析并使用基于 X-means 算法的聚类分析，根据 DGA 域名的字符相似性和查询行为相似性，通过聚类和集合分析方法对疑似恶意域名进一步检测，降低系统误检率。

《面向无载体信息隐藏的映射关系智能搜索方法》提出了一种面向无载体信息隐藏的基于深度学习映射关系智能搜索方法，从已有图像库出发，基于深度神经网络，自动搜索一套高容量、高覆盖率的映射关系，从而解决传统人工方法存在的传输开销大、图像库建立困难的问题。

《基于 HOG-SVM 的跳频信号检测识别算法》提出了一种基于方向梯度直方图与支持向量机的跳频信号检测识别算法，在室内多径信道环境下进行了测试验证，自动化的精确检测到了开放电磁环境下的跳频信号并且实现了对多种跳频序列的识别。

《智能合约安全漏洞研究综述》系统分析了智能合约的特性及其带来的全新安全风险，提出了智能合约安全的三层威胁模型，总结了智能合约安全研究在漏洞方面的进展和挑战，并对智能合约未来安全研究进行了展望，提出了两个潜在的发展方向。

《DC-Hunter：一种基于字节码匹配的危险智能合约检测方案》提出了一种基于字节码匹配的智能合约漏洞检测方案，可以通过已知的漏洞合约找到类似的漏洞合约，并且可以直接应用于现实世界中的智能合约，无需源码和预先定义的漏洞特征。同时揭露了一种新型称为“蜜罐合约”的危险合约。

《小波域基于差分统计量直方图平移的图像鲁棒可逆信息隐藏算法》提出了一种具有更高鲁棒性的算法，为了得到带秘密信息的图像，对载体图像进行 Haar 小波变换，并计算出分块的差分统计量以构造差分统计量的直方图，然后通过可逆的整数变换实现直方图平移，从而将秘密信息嵌入到图像的低频子带中。

我们要特别感谢《信息安全学报》编委会对本期专题工作的信任和指导, 感谢编辑部各位工作人员从征稿启事发布、审稿专家邀请至评审意见汇总、论文定稿、修改、校对和出版所付出的辛勤工作和汗水, 非常感谢专题评审专家及时、专业、细致的评审。我们还要感谢向专题踊跃投稿的各位作者。

最后, 感谢本期专题的读者们, 希望专题能够有助于你们的技术研究工作。