人工智能安全专栏前言

人工智能已成为当今世界科技发展和技术变革的重要驱动力，其在计算机视觉、语音识别和自然语言处理领域的广泛应用，已深刻地影响到人类的日常生活。然而，人工智能面临多种隐私和安全问题，并且人工智能系统的高复杂性和难解释性，导致这些安全隐患无法得到有效的检测和预防。特别是航空航天、智慧医疗和无人驾驶等安全攸关的领域，在安全性、可靠性和可解释性方面对人工智能算法和模型提出了更高的要求。因此，如何保证人工智能的安全成为了当前国内外的研究热点。本专栏旨在推动人工智能安全的学术研究及产业实践，探索人工智能安全与隐私保护的新理论、新方法和新技术，并展示国内研究人员在该领域的最新研究成果。

本专栏涵盖的主题包括人工智能模型的验证和测试、对抗机器学习、人工智能数据隐私和知识产权保护、联邦学习和多方安全计算、人工智能系统的框架和实现安全、隐私保护的数据挖掘、人工智能在信息安全领域的应用、人工智能安全标准与评估检测、人工智能安全对社会和经济的影响。历时近 6 个月的征稿，本专栏共收到了数十篇投稿，每一篇稿件经过多位审稿专家专业、认真和及时的评审，最终有 12 篇文章被专栏收录，其中包括 1 篇综述性文章和 11 篇技术性文章。

在人工智能安全方向，本专栏收录了 5 篇文章，主要涵盖了人工智能安全框架、模型水印、针对人脸检测的对抗攻击以及隐私保护。《人工智能安全框架》为促进人工智能产业健康有序的发展，提出了一套人工智能安全框架，从人工智能安全目标、人工智能安全分级能力以及人工智能安全技术和管理体系 3 个方面阐述了人工智能发展应遵循的安全规范和标准，为提升人工智能产业的安全防护能力提供了重要的参考。《人工智能模型水印研究综述》针对当前人工智能模型的知识产权问题，系统性地调研了模型水印的技术和算法，并复现了部分典型算法进行效果的比较，最后提出在模型水印领域中 5 个具有挖掘潜力的发展方向，对于该领域的研究人员具有很好的指导价值。《针对人脸检测对抗攻击风险的安全测评方法》总结了现行人脸检测模型所面临的安全风险以及物理域对抗攻击的难度，结合集成学习思想，提取多个人脸检测模型的公共注意力热力图，并利用该热力图实施黑盒人脸检测对抗攻击。该攻击方法能够有效地攻破刷脸支付和美颜相机等软件，为物理世界的人工智能模型安全评估提供了重要的参考。《基于特征梯度的调制识别深度网络对抗攻击方法》针对自动调制识别(Automatic Modulation Recognition)的深度神经网络，提出了一种基于特征梯度的对抗攻击方法，在白盒和黑盒两种攻击场景下均可以有效地攻击深度神经网络提取的调制信号空时特征，且生成的对抗样本具有更好的迁移性。差分隐私虽然能够有效提升机器学习模型的隐私保护能力，但带来的数据噪音可能会影响到模型的判别准确率，因此，《基于特征映射的差分隐私保护机器学习方法》提出了一种基于特征映射的差分隐私保护机器学习方法，结合预训练神经网络和影子模型训练技术，以差分向量的形式将原数据样本的特征向量映射到高维向量空间，缩短样本在高维向量空间的距离，以减少模型更新造成的隐私信息泄露，同时提高机器学习模型的隐私保护能力和分类能力。

在人工智能辅助安全方向，本专栏收录了 5 篇文章，内容涵盖渗透测试、入侵检测、流量分类及 SQL 注入检测。《基于深度强化学习的智能化渗透测试路径发现》结合深度强化学习和渗透测试，采用马尔可夫决策模型对渗透测试过程中的状态变迁进行建模，在传统强化学习 DQN 算法的基础上提出了一种 Noisy-Double-Dueling DQNper 改进版本，并在不同规模的网络场景中取得了较快的收敛速度和较高的扩展性。该研究是一项利用深度强化学习改进和增强传统安全解决方案的创新和探索。《DRL-IDS：基于深度强化学习的工业物联网入侵检测系统》针对工业物联网日益增长的网络威胁，设计和实现了一个基于深度强化学习近端策略优化(Proximal Policy Optimization)的入侵检测系统，弥补了传统方法面对新型网络攻击无法自适应地检测、

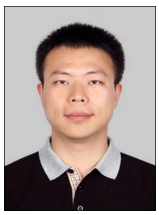
响应和防御等弱点,在大规模的工业物联网真实数据集上取得了良好的检测效果。《对抗攻击威胁基于卷积神经网络的网络流量分类》针对基于卷积神经网络的网络流量分类方法,提出了一种有效的对抗攻击,即对网络流量转换得到的图像添加人眼难以察觉的细微扰动,从而造成卷积神经网络的判别错误。另外,该研究提出了一种基于混合对抗训练的方法来增强分类模型的鲁棒性,以提升对此类对抗攻击的防御能力。《基于变分自编码器的不平衡样本异常流量检测》针对网络流量中正负样本严重不平衡的问题,提出了一种基于变分自编码器的样本生成方法。该方法结合 KNN 和 DBSCAN 算法,筛选出少数类样本与多数类样本接近的样本簇,并利用变分自编码网络模型,对一个或多个子簇中的少类样本进行学习扩充,将扩充后的样本加入到原有样本中以构建新的训练集,最后采用决策树对网络中的异常流量进行识别和判定。《基于信息携带的 SQL 注入攻击检测方法》提出了基于信息携带的 SQL 注入攻击检测方法 SQLIA-IC。该方法利用信息值简化机器学习和标记器来检测样本中的敏感信息,并根据攻击样本携带的信息值进行动态特征匹配,以进行进一步的注入攻击判定,进而提升检测的准确率。

本专栏收录了 2 篇有关深度伪造视频检测的文章。《基于 i_ResNet34 模型和数据增强的深度伪造视频检测方法》为解决深度伪造视频检测中面部特征提取不充分的问题,改进了传统的 ResNet 模型来减小视频帧中人脸面部特征信息的损失,并提出了 3 种基于信息删除的数据增强技术来对人脸数据进行扩充,增加了模型的鲁棒性,为深度视频伪造的检测领域研究提供了研究思路和指导性建议。同样,针对深度视频伪造的问题,《基于 3D CNNS 的深度伪造视频篡改检测》则基于伪造视频的时域特征和空域特征的不一致性,利用 3D CNNS 来训练获得伪造视频检测模型。该方法在检测准确率等方面优于现有的深度伪造视频检测模型。

作为本专栏的特邀编审,我们首先感谢《计算机科学》编委会对本专栏工作的大力支持和指导,感谢《计算机科学》编辑部在论文征稿、审稿、定稿和出版过程中付出的努力,感谢本专栏的评审专家为本次投稿的论文提供了专业、及时和细致的评审意见,保证了专栏收录的文章都保持非常高的水准,感谢踊跃投稿的各位作者在各自的研究领域辛勤耕耘、默默付出,为中国的科技发展和强盛贡献着自己的力量。最后,衷心地希望本专栏能够给读者带来一些思想的启迪和研究的帮助。

中国科学院信息工程研究所　　陈　恺
中国科学院信息工程研究所　　孟国柱
浙江工商大学　　邵　俊
中国农业大学　　吕春利
北京理工大学　　田东海

专栏特邀编审



陈　恺　中国科学院信息工程研究所研究员，博士，博士生导师，中国科学院大学教授，信息安全国家重点实验室副主任，《信息安全学报》编辑部主任，中国计算机学会系统软件专委会常委。主要研究领域包括软件与系统安全、人工智能安全。在 S&P，USENIX Security，CCS 等高水平会议和期刊发表论文 100 余篇；曾主持国家自然科学基金重点项目等 40 余项；入选国家“万人计划”青年拔尖人才、北京市“杰青”、北京市智源青年科学家等。

孟国柱　中国科学院信息工程研究所副研究员，硕士生导师。主要从事软件与系统安全、人工智能安全与隐私的教学和科研工作。在 USENIX Security，ICSE，FSE 等国际高水平学术会议和期刊发表论文 30 余篇；获得了 2020 年天津市科技进步一等奖、2019 ACM SIGSAC 中国科技新星、ACM SIGSOFT ICSE 2018 最佳论文等多项奖励；承担并参与了国家科技部重大专项、国家自然科学基金委青年项目、CCF-腾讯犀牛鸟科研基金等近 10 项科研项目。

邵　俊　浙江工商大学计算机与信息工程学院教授，博士毕业于上海交通大学，获上海市优秀博士学位论文奖。研究领域包括应用密码学、区块链技术、云/雾计算安全、人工智能安全等。在 CCF 推荐会议和期刊上发表论文数篇，包括 IEEE TIFS，IEEE TDSC，INFOCOM，ESORICS，PKC 等；担任多个 CCF 推荐会议程序委员会委员，如 ACISP，ICC 等；入选“浙江省高校领军人才培养计划”和浙江省“151 人才工程”等；主持了浙江省自然科学基金杰出青年项目。

吕春利　中国农业大学副教授，数据科学与工程系副主任，中国农业大学计算中心副主任，北京市青年教学名师。主要研究方向为信息安全、数据处理与分析、区块链、人工智能等。主持并参加了国家科技重大专项、国家自然科学基金、公安部三所课题 10 余项；在国内外学术会议和期刊上发表论文 10 余篇，授权发明专利 8 项。

田东海　北京理工大学计算机学院副教授，硕士生导师。2012 年 3 月博士毕业于北京理工大学，获工学博士学位；2009 年 8 月至 2011 年 8 月受国家留学基金委资助赴美国宾夕法尼亚州立大学进行联合培养；2012 年 3 月毕业留校后先后在软件学院和计算机学院工作。主要研究方向为软件安全、人工智能安全、恶意软件分析、智能终端安全、云计算安全等。近年来作为项目负责人主持了 8 项纵向科研课题，包括：国家自然科学青年基金、CCF-绿盟科技“鲲鹏”科研基金、信息安全国家重点实验室开放课题、中科院网络评测技术重点实验室开放课题(2 项)、上海市信息安全综合管理技术研究重点实验室开放课题和北京理工大学基础科研基金(2 项)。作为软件安全工程技术北京市重点实验室骨干成员参与了国家重点研发计划、国防基础研究计划等多个国家级和部级科研项目。目前以第一作者(或通信作者)在 IEEE Transactions on Cloud Computing，Software：Practice and Experience，Computer Networks，Future Generation Computer Systems，Expert Systems With Applications，NDSS，ISC，SecureComm 等国内外重要期刊和会议上发表论文 30 余篇，其总引用数超过 230 次，授权国家发明专利 3 项，出版学术专著 1 本。担任多个国际学术会议程序委员会委员，担任 IET Information Security，Computer Networks，Expert Systems With Applications，Computers & Electrical Engineering 等多个重要 SCI 期刊的审稿人。