安全管理部
Security Management Department

赛博研究院

WORLD ARTIFICIAL INTELLIGENCE CONFERENCE
2018世界人工智能大会

人工智能赋能网络空间安全

模式与实践

安全管理部
Security Management Department

赛博研究院

版权声明

出品方：

腾讯公司安全管理部

上海赛博网络安全产业创新研究院

撰稿人：

韩李云 腾讯生态安全中心高级研究员

赵文俊 腾讯反诈骗防控中心安全专家

杨　乐 腾讯生态安全中心研究员

杨晓光 腾讯安全管理部安全标准专家

惠志斌 上海社会科学院互联网研究中心执行主任

石建兵 上海赛博网络安全产业创新研究院高级研究员

李　宁 上海赛博网络安全产业创新研究院高级研究员

咨询专家:

徐英瑾 复旦大学哲学学院教授

刘山泉 上海经信委系统安全处处长

王　琦 碁震KEEN公司创始人兼CEO、GeekPwn大赛创办人

袁劲松 斗象科技创始人

于　旸 腾讯玄武实验室总监

吕一平 腾讯科恩实验室总监

鞠　奇 腾讯信息安全部技术总监

曹建峰 腾讯研究院法律研究中心高级研究员

韩昕辉 腾讯手机管家安全专家

崔培豪 腾讯安全策略专家

本报告中技术应用和案例部分得到腾讯公司守护者计划安全团队、腾讯安全联合实验室、信息安全部、微信安全风控中心、优图实验室、网络安全与犯罪基地、政务舆情部等多个安全团队的鼎力支持，特别鸣谢腾讯安全管理部刘鑫、史波良、周正、陈晓冉、刘澎、门美子、翟尤、赵玉现、武杨、黄超等诸位专家、研究员对本报告编写给予的大力支持和宝贵建议。

安全管理部 Security Management Department　SIGSI 赛博研究院

安全管理部 Security Management Department　SIGSI 赛博研究院

摘　要

人工智能时代，网络空间安全威胁全面泛化，如何利用人工智能思想和技术应对各类安全威胁，是国内外产业界共同努力的方向。本报告从风险演进和技术逻辑的角度，将网络空间安全分为网络系统安全、网络内容安全和物理网络系统安全三大领域；在此基础上，本报告借鉴Gartner公司的ASA自适应安全架构模型，从预测、防御、检测、响应四个维度，提出人工智能技术在网络空间安全领域的具体应用模式。与此同时，本报告结合国内外企业最佳实践，详细阐释人工智能赋能网络空间安全（AI+安全）的最新进展。最后，本报告提出，人工智能安全将成为人工智能产业发展最大蓝海，人工智能的本体安全决定安全应用的发展进程，“人工”+“智能”将长期主导安全实践，人工智能技术路线丰富将改善安全困境，网络空间安全将驱动人工智能国际合作。

安全管理部 Security Management Department 赛博研究院

目　录

安全管理部 Security Management Department 赛博研究院

第一章 人工智能技术的发展沿革

人工智能的起源几乎与电子计算机技术发展同步，早在20世纪40年代，来自数学、心理学、工程学等多领域的科学家就开启了通过机器模拟人类思想决策的科学研究，被誉为“人工智能之父”的阿兰·图灵在1950年提出了检验机器智能的图灵测试，预言了智能机器的出现和判断标准。1956年，美国达特茅斯会议正式诞生“人工智能”（Artificial Intelligence）的概念。此后的60多年里，伴随信息技术以及生物、机械等相关技术发展，人工智能在概念和技术上不断扩展和演进。对于何谓人工智能，人们尚未达成完全的共识，但从较为流行的学派和定义中可以从两个方面去认识和理解它，一方面，人工智能是“像人一样思考”的系统，解决的是逻辑、推理和寻找最优解等问题，比如认知架构和神经元网络；另一方面，人工智能是“像人一样行动”的系统，可以通过认知、计划、推理、学习、沟通、决策等行动实现目标，比如智能软件代理、机器人。

（一）人工智能技术的关键阶段

人工智能概念自首次提出以来，经历了长期而又波折的算法演进和应用检验，总的来说，其早期发展阶段较为平缓，也曾几度陷入低谷，直至近20年随着计算技术和大数据技术的高速发展，人工智能得到超强算力和海量数据的支持，获得越来越广泛的应用验证。目前的人工智能处于加速发展期，无论是技术本身的演进还是应用领域的扩展，都取得了跨越式发展。

纵观人工智能发展历程，大致可分为三个阶段：

模式识别（Pattern Recognition）阶段：最初的模式识别阶段大致从20世纪50年代前后延续至20世纪80年代，此时期的人工智能技术主要集中在模式识别类技术的研发和应用上，包括沿用至今的语音识别和图像识别技术均发轫于此。模式识别主要是指模仿人类识读符号的认知过程从而实现智能系统。

机器学习（Machine Learning）阶段：机器学习最早可追溯至人工智能概念诞生不久后，但实际取得突破性进展是在20世纪80年代及以后。彼时的人工智能以应用仿生学为主要特点，受到人脑学习知识主要是通过神经元间突触形成与变化

的启发，人们发现计算机也可用来模拟神经元工作，因此也称为神经元发展阶段，今天广泛应用的人工神经网络（ANN，Artificial Neural Networks）、支持向量机(SVM, Support Vector Machine)技术均来源于此，SVM作为这一时期的最顶峰成果，它实现了高效的归纳学习，具有在数据样本有限情况下精确分类的优势。

深度学习（Deep Learning）阶段：2006年，随着深度学习模型的提出，人工智能引入了层次化学习的概念，通过构建较简单的概念来学习更深、更复杂的概念，真正意义上实现了自我训练的机器学习。深度学习可从大数据中发现复杂结构，具有强大的推理能力和极高的灵活性，由此揭开了崭新人工智能时代的序幕。在人工智能第三波发展热潮中，深度学习逐渐实现了在机器视觉、语音识别、机器翻译等多个领域的普遍应用，也催生了强化学习、迁移学习、生成对抗网络等新型算法和技术方向。

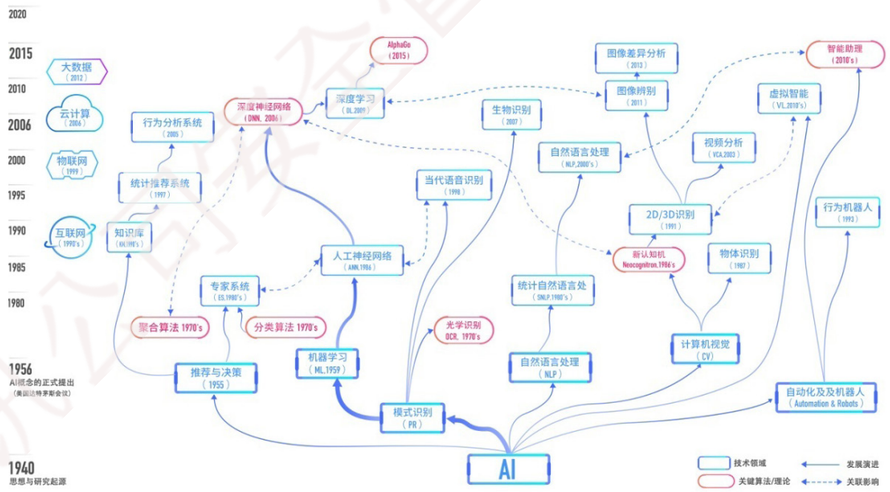


图1人工智能技术发展演进图

（二）人工智能技术的驱动因素

除了计算机技术和网络技术的推动，人工智能的快速发展和广泛应用，还得益于算力提升、数据爆增、算法优化、万物连接等因素的驱动：

- 前所未有的计算能力：人工智能算法模型的训练和计算需耗费大量的时间和计算能力，当前计算机和芯片等硬件水平的提升，加上计算机网络技术发展推动云计算架构的弹性部署等，为人工智能训练更为复杂的模型提供了前所未有的计算能力。
- 海量的可用训练数据：随着互联网大规模服务集群产生、搜索和电商业务蓬勃发展产生大量的数据积累，从根本上促进了深度学习的诞生和人工智能的突破性发展。现阶段，不仅数据所承载的信息量增大，信息形态也越发多元，从文本、声音、图像等原始数据发展到行为数据、环境数据等，数据数量和质量综合影响着工智能算法的“人”性和思考水平，为人工智能普遍应用提供了可能。
- 更成熟的人工智能算法：以深度学习（DL）为代表的算法不断精进，极速融合各应用场景，这类技术基于大量数据分析和训练进行自主学习，实际应用成果更符合人类认知，因此被广泛接受。
- 广域异构的网络连接：当前全球已基本实现有线、无线等方式繁密而充分的网络连接，使人与机器、机器与机器的连接更为普遍，信息和数据以空前的方式和速度流动、汇聚、共享，这种连接与融合趋势不断重塑人与机器的关系，成为人工智能技术应用的极佳场景。

（三）人工智能技术的典型代表

- 机器学习(ML, Machine Learning):是当前人工智能的关键技术，通过设定模型，输入数据进行训练，改善自身性能，重在归纳、聚合而非演绎。
- 专家系统（ES，Expert System）：专家系统主要是将规则和逻辑引入AI系统，帮助和执行自动化决策。
- 过程自动化(AT,Automation)：采用自动化脚本的方法，实现任务自动化代替或协助人类员工。
- 深度学习（DL，Deep Learning）：深度学习是机器学习中一种基于对数据进行表征学习的方法，使用特定的表示方法从实例中更容易学习新的任务。
- 自然语言处理（NLP, Natural Language Processing）：让计算机处理并

理解人类所使用的各类语言，广义定义还包含让计算机正确运用人类语言自如地与人进行多种形式的沟通。

- 计算机视觉(CV, Computer Vision)：研究让计算机如何“看”世界，常见的有对图像进行分析的图像处理技术（IP，Image Processing）、从动态视频获取有效信息的视频分析技术(VA，Video Analysis)等，还有支持AR和VR等的虚拟智能技术（VI, Virtual Intelligence）等新兴技术
- 模式识别(PR，Pattern Recognition)：对信息进行整合与智能分析，对由环境和客体组成的模式进行自动处理和判读，技术实现可分为有监督的分类（Supervised Classification）和无监督的分类(Unsupervised Classification)两种。
- 情绪识别（ER，Emotion Recognition）：综合多种技术感知人类的情绪状态。
- AI建模（DT，Digital Twin/AI Modeling）：通过软件来沟通物理系统与数字世界，这也是物理与虚拟世界的交界面。
- 机器人技术(RB，Robotics):机器人有着广阔的应用，形态也各异，常见的有无人驾驶、无人机等。
- 虚拟代理(VA, Virtual Agents)：复合多项技术，能够与人类进行交互的计算机代理或程序，目前常被用于客户服务或语音助理。

（四）人工智能技术的广泛应用

人工智能技术本身是一项具有巨大变革潜力的科技进步，人工智能在越来越多领域的普及应用，对整个经济体系产生了深远影响。机器人生产在基础制造业上的优势已经逐渐获得了整个行业的认同，随着无人驾驶、智能决策系统等新技术逐渐成熟，金融服务、仓储行业、交通运输、城市管理、甚至法律、教育和医疗等行业，都将被深深卷入人工智能所引发的技术变革浪潮。这是一场具有产业革命意义的技术变革，它有可能深远地改变人类生产的基本形态，也将深刻地改变当代社会与经济的结构。

具体而言，人工智能在各个领域的应用包括：在医疗领域，人工智能可用于医疗影像诊断、药物研发、虚拟医生助手、可穿戴设备等；在金融领域，人工智能在智能投顾、金融风险控制、金融预测与反欺诈、智慧理财等方面均已崭露头角；在教育领域，人工智能可用于智能评测、个性化辅导、儿童陪伴等；在电商领域，人工智能可用于仓储物流、智能导购和客服；在智慧家居领域，人工智能可用于智能家电、智能家具、家庭管家、陪护机器人等领域。此外，人工智能在智能制造、智能交通、自动驾驶、智慧城市、智慧安防等领域不断引领人类社会的快速变迁。

第二章 网络空间安全的内涵与态势

（一）网络空间安全的内涵

网络空间的概念是指由现代信息技术革命产生的，由通信线路和设备、计算机、软件、数据、用户以及任何接入网络的物体等要素交互形成的全新空间，涵盖物理设施、用户和内容逻辑等多个层面，它将生物、物体和自然空间之间建立起智能联系，是人类社会活动和财富创造的全新领域。

网络空间安全是网络空间中所有要素和活动免受来自各种威胁的状态。随着信息技术的不断创新发展，网络空间安全的范畴正不断扩大，成为非传统安全的重要组成部分，并与国家、政治、社会、经济领域的安全密不可分。

从网络信息技术的发展历程以及技术逻辑，网络空间安全可分为三大领域，分别为网络系统安全、网络内容安全和物理网络系统安全。其中，网络系统安全包括信息基础设施、计算机系统、网络连接、用户数据等设备和信息的安全保障，需要抵御各种恶意攻击对信息和网络系统的入侵、渗透、中断、破坏，以及对用户数据的泄露、窃取，网络系统安全是保障全球网络和计算机系统稳定运行，保护用户数据和隐私的基础。网络内容安全是指在网络环境中产生和流转的信息内容是否合法、准确和健康，是否会对政治、经济、社会和文化产生不良影响和危害。物理网络系统安全包括网络空间中任何与网络连接的物、人等物理要素的安全，随着物联网、脑机接口、机器人等技术的迅猛发展，网络空间的威胁已延伸到物理空间和现实世界，由此产生对资产、人身以及自然环境等要素的潜在安全威胁。网络系统安全、网络内容安全和物理网络系统安全相互影响和融合交织，构成了本报告网络空间安全的基本内涵。

（二）人工智能时代网络空间安全发展态势

1、网络空间安全威胁趋向智能

随着网络信息技术全面普及以及数据价值的持续增长，网络空间安全威胁更趋严峻，且呈现出智能化、隐匿性、规模化的特点，网络空间安全的防御、检测和响应面临更大的挑战。采用人工智能的网络威胁手段已经被广泛应用于网络犯罪，包括漏洞自动挖掘、恶意软件智能生成、智能化网络攻击等，网络攻击方式的智能化升级打破了攻防两端的平衡。魔高一尺道高一丈，网络安全攻防不对称要求网络空间安全防御方采取更加智能化的思想与手段予以应对。

2、网络空间安全边界开放扩张

智能互联时代，网络空间安全的边界不断扩展。一方面，传统基于网络系统和设备等物理边界的网络安全防御边界日趋泛化，网络安全攻击范围被全面打开。另一方面，网络空间治理渗透在政治、经济、社会等各个领域，网络空间安全影响领域全面泛化。边界的开放扩张要求积极将各类智能化技术应用于全业务流程的安全防御。

3、网络空间安全人力面临不足

网络空间安全威胁形势日趋严峻，与之对应的是安全人员面临严重短缺，根据迈克菲调查显示，企业普遍认为他们需要增加24%的安全人员才能有效应对面临的网络威胁。网络空间不断延展、移动设备增加、多云端服务正在使安全人员的工作变得越来越复杂，而安全人员的短缺更是加剧了安全风险问题。利用人工智能等技术推动网络防御系统的自主性和自动化，降低安全人员风险分析和处理压力，辅助其更加高效地进行网络安全运维与监控迫在眉睫。

4、网络空间安全防御趋向主动

针对层出不穷、花样翻新、破坏加剧的恶意代码、漏洞后门、拒绝服务攻击、APT攻击等安全威胁，现有被动防御的安全策略显得力不从心。智能时代，网络空间安全从被动防御趋向主动防御，人工智能驱动的自动化防御能够更快更好地识别威胁，缩短响应时间，是网络空间安全发展的必然方向和破解之道。

第三章 人工智能在网络空间安全领域的应用模式

人工智能技术日趋成熟，人工智能在网络空间安全领域的应用（简称AI+安全）不仅能够全面提高网络空间各类威胁的响应和应对速度，而且能够全面提高风险防范的预见性和准确性。因此，人工智能技术已经被全面应用于网络空间安全领域，在应对智能时代人类各类安全难题中发挥着巨大潜力。

（一）AI+安全的应用优势

人们应对和解决安全威胁，从感知和意识到不安全的状态开始，通过经验知识加以分析，针对威胁形态做出决策，选择最优的行动脱离不安全状态。类人的人工智能，正是令机器学会从认识物理世界到自主决策的过程，其内在逻辑是通过数据输入理解世界，或通过传感器感知环境，然后运用模式识别实现数据的分类、聚类、回归等分析，并据此做出最优的决策推荐。当人工智能运用到安全领域，机器自动化和机器学习技术能有效且高效地帮助人类预测、感知和识别安全风险，快速检测定位危险来源，分析安全问题产生的原因和危害方式，综合智慧大脑的知识库判断并选择最优策略，采取缓解措施或抵抗威胁，甚至提供进一步缓解和修复的建议。这个过程不仅将人们从繁重、耗时、复杂的任务中解放出来，且面对不断变化的风险环境、异常的攻击威胁形态比人更快、更准确，综合分析的灵活性和效率也更高。

因此，人工智能的“思考和行动”逻辑与安全防护的逻辑从本质上是自洽的，网络空间安全天然是人工智能技术大显身手的领域。

（1）基于大数据分析的高效威胁识别：大数据为机器学习和深度学习算法提供源源动能，使人工智能保持良好的自我学习能力，升级的安全分析引擎，具有动态适应各种不确定环境的能力，有助于更好地针对大量模糊、非线性、异构数据做出因地制宜的聚合、分类、序列化等分析处理，甚至实现了对行为及动因的分析，大幅提升检测、识别已知和未知网络空间安全威胁的效率，升级精准度和自动化程度。

（2）基于深度学习的精准关联分析：人工智能的深度学习算法在发掘海量数据中的复杂关联方面表现突出，擅长综合定量分析相关安全性，有助于全面感知内外部安全威胁。人工智能技术对各种网络安全要素和百千级维度的安全风险数据进行归并融合、关联分析，再经过深度学习的综合理解、评估后对安全威胁的发展趋势做出预测，还能够自主设立安全基线达到精细度量网络安全性的效果，从而构建立体、动态、精准和自适应的网络安全威胁态势感知体系。

（3）基于自主优化的快速应急响应：人工智能展现出强大的学习、思考和进化能力，能够从容应对未知、变化、激增的攻击行为，并结合当前威胁情报和现有安全策略形成适应性极高的安全智慧，主动快速选择调整安全防护策略，并付诸实施，最终帮助构建全面感知、适应协同、智能防护、优化演进的主动安全防御体系。

（4）基于进化赋能的良善广域治理：随着网络空间内涵外延的不断扩展，人类面临的安全威胁无论从数量、来源、形态、程度和修复性上都在超出原本行之有效的分工和应对能力，有可能处于失控边缘，人工智能对人的最高智慧的极限探索，也将拓展网络治理的理念和方式，实现安全治理的突破性创新。人工智能不仅能解决当下的安全难题，而通过在安全场景的深化应用和检验，发现人工智能的缺陷和不足，为下一阶段的人工智能发展和应用奠定基础，指明方向，推动人工智能技术的持续变革及其更广域的赋能。

（二）AI+安全的产业格局

人工智能以其独特的优势正在各类安全场景中形成多种多样的解决方案。从可观察的市场指标来看，近几年来人工智能安全市场迅速成长，IDC公司在2017年就估算，到2020年仅人工智能及认知科学领域的市场规模将达到470亿美元。而MarketResearch公司在2018年的研究表明，在网络安全中人工智能应用场景增多，同时地域覆盖范围扩大，将进一步扩大技术在安全领域的应用，因此人工智能技术在安全市场内将快速发展，预计到2024年，可用在安全中的人工智能技术市场规模将超过350亿美元，在2017-2024年之间年复合增长率（CAGR）可达31%。MarketsandMarkets公司在2018年1月发布的《安全市场中人工智能》报告则认为，2016年AI安全市场规模就已达29.9亿美元、2017年更是达到39.2亿

美元，预测在2025年将达到348.1亿美元，年复合增长率为31.38%。而爱尔兰的Research and Markets公司在2018年4月份发布了专门的市场研究报告，认为到2023年人工智能在安全领域应用的市场规模将达182亿美元，年复合增长率为34.5%。由于机器学习对付网络犯罪较为有效，因此机器学习作为单一技术将占领最大的一块市场，到2023年其市场规模预计可达60亿美元。

除了传统安全公司致力于人工智能安全，大型互联网企业也在积极开展人工智能安全实践，如Google、Facebook、Amazon、腾讯、阿里巴巴等均在围绕自身业务积极布局人工智能安全应用。

（三）AI+安全的实现模式

人工智能是以计算机科学为基础的综合交叉学科，涉及技术领域众多、应用范畴广泛，其知识、技术体系实际与整个科学体系的演化和发展密切相关。因此，如何根据各类场景安全需求的变化，进行AI技术的系统化配置尤为关键。

本报告采用Gartner公司2014年提出的自适应安全架构（ASA，Adaptive Security Architecture）来分析安全场景中人工智能技术的应用需求，此架构重在持续监控和行为分析，统合安全中预测、防御、检测、响应四层面，直观的采用四象限图来进行安全建模。其中“预测”指检测安全威胁行动的能力；“防御”表示现有预防攻击的产品和流程；“检测”用以发现、监测、确认及遏制攻击行为的手段；“响应”用来描述调查、修复问题的能力。

本报告将AI+安全的实现模式按照阶段进行分类和总结，识别各领域的外在和潜在的安全需求，采用ASA分析应用场景的安全需求及技术要求，结合算法和模型的多维度分析,寻找AI+安全实现模式与适应条件，揭示技术如何响应和满足安全需求，促进业务系统实现持续的自我进化、自我调整，最终动态适应网络空间不断变化的各类安全威胁。

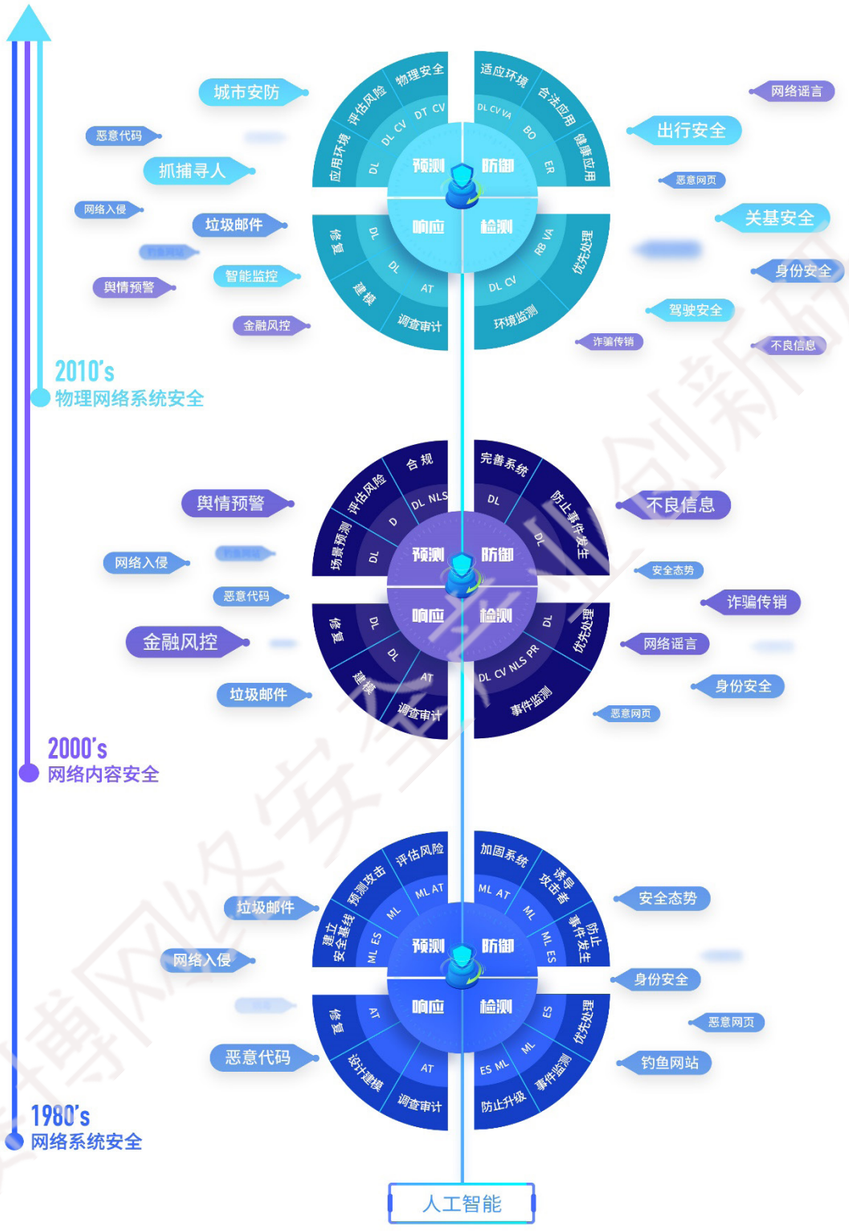
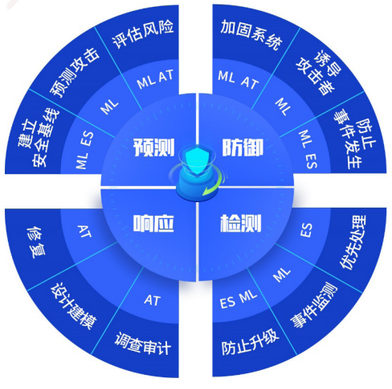


图2人工智能技术在网络空间安全实现模式示意

1、人工智能应用于网络系统安全

人工智能技术较早应用于网络系统安全领域，从机器学习、专家系统以及过程自动化等到如今的深度学习，越来越多的人工智能技术被证实能有效增强网络系统安全防御：

- 机器学习(ML, Machine Learning):在安全中使用机器学习技术可增强系统的预测能力，动态防御攻击，提升安全事件响应能力。
- 专家系统（ES，Expert System）：可用于安全事件发生时为人提供决策辅助或部分自主决策。
- 过程自动化(AT， Automation)：在安全领域中应用较为普遍，代替或协助人类进行检测或修复，尤其是安全事件的审计、取证，有不可替代的作用。
- 深度学习（DL，Deep Learning）：在安全领域中应用非常广泛，如探测与防御、威胁情报感知，结合其他技术的发展取得极高的成就。



网络系统安全

图3 人工智能在网络系统安全中的实现模式

如图3所示，通过分析人工智能技术应用于网络系统安全，在四个层面均可有效提升安全效能：

预测：基于无监督学习、可持续训练的机器学习技术，可以提前研判网络威胁，用专家系统、机器学习和过程自动化技术来进行风险评估并建立安全基线，可以让系统固若金汤。

防御：发现系统潜在风险或漏洞后，可采用过程自动化技术进行加固。安全事件发生时，机器学习还能通过模拟来诱导攻击者，保护更有价值的数字资产，避免系统遭受攻击。

检测：组合机器学习、专家系统等工具连续监控流量，可以识别攻击模式，实现实时、无人参与的网络分析，洞察系统的安全态势，动态灵活调整系统安全策略，让系统适应不断变化的安全环境。

响应：系统可及时将威胁分析和分类，实现自动或有人介入响应，为后续恢复正常并审计事件提供帮助和指引。

因此人工智能技术应用于网络系统安全，正在改变当前安全态势，可让系统弹性应对日益细化的网络攻击。在安全领域使用人工智能技术也会带来一些新问题，不仅有人工智能技术用于网络攻击等伴生问题，还有如隐私保护等道德伦理问题，因此还需要多种措施保证其合理应用。总而言之，利用机器的智慧和力量来支持和保障网络系统安全行之有效。

2、人工智能应用于网络内容安全

人工智能技术可被应用于网络内容安全领域，参与网络文本内容检测与分类、视频和图片内容识别、语音内容检测等事务，切实高效地协助人类进行内容分类和管理。面对包括视频、图片、文字等实时海量的信息内容，人工方式开展网络内容治理已经捉襟见肘，人工智能技术在网络内容治理层面已然不可替代。

在网络内容安全领域所应用的人工智能技术如下：

- 自然语言处理（NLP，Natural Language Processing）：可用于理解文字、语音等人类创造的内容，在内容安全领域不可或缺。
- 图像处理（IP，Image Processing）：对图像进行分析，进行内容的识别和分类，在内容安全中常用于不良信息处理。

- 视频分析技术(VA, Video Analysis)：对目标行为的视频进行分析，识别出视频中活动的目标及相应的内涵，用于不良信息识别。

网络内容安全

图4 人工智能在网络内容安全中的实现模式

如图4所示，通过分析人工智能技术应用于网络内容安全，在四个层面均可有效提升安全效能：

预防阶段：内容安全最重要的是合规性，由于各领域的监管法律/政策的侧重点不同而有所区别且动态变化。在预防阶段，可使用深度学习和自然语言处理进行相关法律法规条文的理解和解读，并设定内容安全基线，再由深度学习工具进行场景预测和风险评估，并及时将结果向网络内容管理人员报告。

防御阶段：应用深度学习等工具可完善系统，防范潜在安全事件的发生。

检测阶段：自然语言、图像、视频分析等智能工具能快速识别内容，动态比对安全基线，及时将分析结果交付给人类伙伴进行后续处置，除此之外，基于内容分析的情感人工智能也已逐步应用于舆情预警，取得不俗成果。

响应阶段：在后续调查或留存审计资料阶段，过程自动化同样不可或缺。

3、人工智能应用于物理网络系统安全

随着物联网、工业互联网、5G等技术的成熟，网络空间发生深刻变化，人、物、物理空间通过各类系统实现无缝连接，由于涉及的领域众多同时接入的设备数量巨大，传感器网络所产生的数据可能是高频低密度数据，人工已经难以应对，采用人工智能势在必行。但由于应用场景极为复杂多样，可供应用的人工智能技术将更加广泛，并会驱动人工智能技术自身新发展。

- 情绪识别（ER，Emotion Recognition）：不仅可用图像处理或音频数据获得人类的情绪状态，还可以通过文本分析、心率、脑电波等方式感知人类的情绪状态，在物理网络中将应用较为普遍，通过识别人类的情绪状态从而可与周边环境的互动更为安全。
- AI建模（DT，Digital Twin/AI Modeling）：通过软件来沟通物理系统与数字世界。
- 生物特征识别(BO, Biometrics)：可通过获取和分析人体的生理和行为特征来实现人类唯一身份的智能和自动鉴别，包括人脸识别、虹膜识别、指纹识别、掌纹识别等技术。
- 虚拟代理(VA, Virtual Agents)：这类具有人类行为和思考特征的智能程序，协助人类识别安全风险因素，让人类在物理网络世界中更安全。

图5 人工智能在物理网络系统安全中的实现模式

物理网络安全由于应用领域广、层次多，可应用的技术类型也极为复杂，因此需要以人为中心，通过全程监测人和系统、人与机器、人与环境之间的交互，确保人与物的不受威胁。以物联网为例，在业务运营系统与网络系统进行融合后，通常可分为负责业务信息的信息技术（IT）网络与负责生产运行维护的运营技术（OT）网络这两部分，IT部分的AI应用与网络系统安全需求基本一致，而OT部分则涉及业务运营安全，与应用场景融合的安全需求变得复杂，安全不仅关乎这些IT资产拥有者的安全，而且其中部分属于关键基础设施，一旦出现风险则将可能给国计民生带来恶劣影响。但人工智能应用得当，不仅可以更高效的抵御OT风险，还可以提升OT运营效能，从而直接创造价值，因此AI应用于OT安全领域不仅在安全管理上成为必需，而且也能促进综合效益提升。例如，AI应用于智慧城市的智能交通安全中的流量管控中，一般通过历史数据与深度学习进行交通流量预测，可通过交通设施的优化布局进行流量管控预防，而在当前道路体系中应用专家系统进行预防，同时通过计算机图像分析技术首先进行道口流量实时监控、车辆通过监测，提前疏解交通堵点，洞悉交通现场态势，即便发生事故，也可迅速调集应急资源快速进行现场处置，在事件回溯期间，可将相关系统中已处理的信息进行自动关联，从而为后续调整道路通行方案提供依据。

第四章 人工智能在网络空间安全领域的应用案例

网络系统安全篇

（一）病毒及恶意代码检测与防御

1、安全需求

传统反病毒查杀流程主要是先通过主动或被动方式获取病毒样本，了解其运行机制与作案原理，然后人工提取特征码或设定病毒查杀策略，再输入反病毒引擎对病毒实施查杀。由此可见，传统病毒查杀过程中人工判断、分析、处理环节所占比例较大，使病毒查杀陷入恶意病毒肆意增长、人工技术分析能力滞后的困局。

2、AI+

人工智能技术的迅猛发展给反病毒引擎的更新升级提供了更多技术支持和创新思路，有机结合了传统病毒查杀经验与人工智能技术打造的新型智能化反病毒引擎已实现应用落地。一方面，在终端设备上加载可支持运行AI算力的芯片以此提升硬件水平；另一方面，基于机器深度学习的下一代反病毒引擎可保持持续的自学习、自适应能力，自动化、智能化地跟进病毒的行为演进，增进对病毒行为的预测、识别和阻断能力。

针对Android系统的移动终端，智能反病毒引擎可深入应用框架层（Framwork）对应用的敏感行为进行监控。基于AI芯片的计算能力来运行AI反病毒模型，可实现独立、安全地检测判断应用是否为恶意，而在前端通过安全应用检测的结果展示和用户授权交互，及时阻断恶意行为，卸载恶意应用，为用户提供实时安全防护。

3、案例实践

目前，多家互联网公司都在智能反病毒引擎产品的开发上取得重大成果。例如，腾讯安全团队基于真实运行行为、系统层监控、AI芯片检测、AI模型云端训

练、神经网络算法自主研发了腾讯TRP-AI反病毒引擎，该引擎相较于传统引擎具有抗免杀、高性能、实时防护、可检测0Day病毒等优势，引擎可自动化训练，大大缩小了查杀周期和运营成本。据《腾讯TRP-AI反病毒引擎白皮书》数据显示，通过使用TRP-AI反病毒引擎，可使病毒检测覆盖率达到90%，检测准确率高达99%，病毒发现能力提升8%，发现速度提升12%，病毒发现快于病毒传播的占比提升到92%，检测耗时30ms远低于传统引擎的100~200ms，平均性能消耗保障对设备使用体验零影响。

此外，SparkCognition、Cylance、Deep Instinct 和Invincea等国外公司也将人工智能技术运用在检测恶意软件这一领域。比如SparkCognition打造的“认知”防病毒系统DeepArmor，正是利用了人工智能技术发现和掌握新型恶意软件攻击行为，识别通过变异尝试绕过安全系统的病毒，可准确发现和删除恶意文件，保护网络免受未知病毒和恶意代码威胁。

（二）网络入侵检测与防御

1、安全需求

防火墙和网络入侵检测是计算机网络系统的两把保护伞。防火墙因本身存在的漏洞和后门，不能阻止内部攻击和难以针对一些可以绕开防火墙的外部访问提供实时的入侵检测能力，使网络入侵拦截依然存在漏网之鱼。近年来，蠕虫病毒攻击、DDoS攻击以及攻击漏洞等系统入侵愈演愈烈，大规模、高频率、新技术网络入侵威胁使网络信息系统的机密性、完整性和可用性遭到惨重破坏，网络入侵检测能力的提升迫在眉睫。

2、AI+

结合当下快速发展的人工智能技术研发的新检测技术在检测速度、检测范围和体系结构方面较之传统的入侵检测技术均有大幅优化。入侵检测中利用机器学习算法构建高效准确的分类器和学习器来自主学习入侵行为，特征化入侵规则，从而达到及时准确识别入侵的目的。在众多的机器学习算法中，神经网络、遗传算法、支持向量机、免疫算法、Agent系统在入侵检测中的应用各有所长，比如贝叶斯算法擅长处理不完整和带有噪声的数据集，人工神经网络的抽象能力、学习和自适应能力,长于自我训练并预测未知入侵行为，支持向量机算法的最大优势则是针对有限样本得出最优值，对于数据分类来说是种较为快速的算法， 较好满足入侵检测实时性的要求，而Agent系统算法则凭借其多Agent“局部运作，全局共享”的特点常被应用于分布式入侵检测中。

3、案例实践

基于人工智能技术具有先进性、高效性、准确性等特点，将AI应用于网络入侵检测系统的初创企业逐渐增多，其中Vectra Networks、DarkTrace、Exabeam、CyberX 和BluVector公司表现不俗。校企结合的技术创新模式也取得颇多成果，如2016年麻省理工学院计算机科学和人工智能实验室(CSAIL)与人工智能初创企业PatternEx联合开发的网络安全平台AI2，可以分析挖掘360亿条与安全相关的数据，在数据分析过程中不断训练产生新模型快速改善预测率，从而实现预测、检测和阻止了85%的网络入侵攻击，准确率是当今同类的自动化网络攻击检测系统的2.92倍。

（三）新型身份认证

1、安全需求

身份认证作为确定用户具有的资源访问、使用权限的技术手法，对保证系统和数据安全、维护用户合法权益具有重要意义。当前，我们面临撞库、暴力破解、社工攻击等黑客技术的不断升级和频繁使用，甚至利用人工智能、区块链等新技术破解身份验证系统的风险，传统的账密、U盾等身份认证方式已经难以满足访问、支付等业务的移动化安全需求和个人信息保护需求，尤其是移动支付市场的爆发使移动支付安全问题凸显，其核心问题就是用户的身份验证。

2、AI+

基于数据挖掘和人工智能技术的新型身份认证综合运用了数字签名、设备指纹、时空码、人脸识别等多项身份认证技术，目前在金融、社保、保险、电商、O2O、直播等行业均有广泛应用。其中，多因子身份认证技术能对用户的认

证行为、业务类型、时间、地点等内容深入挖掘，通过机器学习精准识别和拦截恶意认证行为。结合智能语音能够实现对用户行为更全面立体的建模与画像，如智能声纹技术可通过人声识别人的年龄、体重、身高、面部特征和周围环境等信息，指纹、静脉、虹膜等生物识别技术在人工智能技术的加持下精准度均得到大幅提升。在验证码安全方面，利用神经网络算法能够解决传统验证码识别中人工+光学字符识别（OCR）方法带来的人力消耗大、时间成本高的问题，跳过人工对样本预处理，直接用输入神经网络进行自主学习和运算，使得验证码识别率维持在80%左右的高水平。此外，借助大数据分析和机器学习也可更迅速准确地掌握账号或设备关联的各维度信息，包括信誉和价值等，帮助构建基于身份的价值互联网。当然，人工智能技术也会被用于破解验证码，对此只能通过机器学习、识别和自主对抗加以打击。

3、案例实践

腾讯守护者计划基于长期积累的AI人工智能和神经网络的分析能力，引入多维度的动态验证机制对抗黑产。运用人工智能技术对不同类型的验证码图形样本进行学习、运算，再开发验证码接收与分发模块，完成从输入端到输出端的验证码验证识别，将结果反馈至该分布式AI验证码识别系统进行进一步优化。2017年，守护者计划安全团队利用上述技术协助警方刑事打掉“快啊答题”、“光速打码”这两个黑产团伙，这两个团伙均为国内最大的利用人工智能神经网络破解识别验证码的打码平台。

2018年8月15日，支付宝宣布在未来一年里将普及自助收银+刷脸支付，创新采用3D人脸识别技术用于识别验证身份，通过软硬件结合进行活体检测，可以有效防止利用用户照片、视频或软件模拟冒充用户身份支付，据称其识别率高达99.99%，大大提升了身份验证的准确性和电子支付的安全系数。

（四）垃圾邮件检测

1、安全需求

滥发垃圾邮件是目前互联网上最为突出的网络滥用行为之一。根据Securelist发布的全球垃圾邮件数据显示，2017年恶意垃圾邮件数量占邮件总流量的56.63%。[1]肆意传输的垃圾邮件不仅占用了传输、存储和运算资源，也是病毒的主要携带载体。早期的垃圾邮件处理主要依靠滤器（Spam Filter），原理是基于人工规则的模式匹配方法，此类系统维护成本高，也缺乏灵活性。

2、AI+

人工智能在垃圾邮件检测的应用，一般先是对垃圾邮件信息的文本进行分类并将其数值化表示，然后对垃圾信息进行词汇处理、数据清洗、降维、归一化等一系列的数据预处理，再选择适当的机器学习算法来确定每条消息的向量值，实现对垃圾邮件的检测、识别和拦截。也有创新公司专门开发了机器人（虚拟代理），由其代替人与意图利用垃圾邮件行骗的不法分子聊天，占用其时间。

3、案例实践

新西兰的网络安全组织Netsafe研发出一款人工智能机器人（Re:scam，即“回复骗子”），它们能够识别出垃圾邮件并自动回复，以不同年龄、声音、性别、性格模仿人类与骗子聊天，永无止境地向行骗者发问或开玩笑调侃对方，达到拉锯中消耗对方时间精力目的。据统计，“Re:scam”发送给骗子的超过16000余封电子邮件累计“浪费”了骗子们25天以上的时间。

谷歌公司利用人工智能技术在几年前就将Gmail电子邮件服务的垃圾邮件拦截率提升至99.9%、误报率降低至0.05%。Gmail目前拥有高达9亿的全球用户，面对不同地区用户，Gmail垃圾邮件智能过滤器不仅仅是通过预设规则清除垃圾邮件，在运行期间还可根据具体情况自行制定新规则。

（五）基于URL的恶意网页识别

1、安全需求

网络攻击者们常通过钓鱼网站、垃圾广告和恶意软件推广等方式引诱或欺骗不知情的用户访问攻击者提供的恶意网页地址，非法牟利。据国际反钓鱼工作

1 数据来源于：Securelist：2017年Q3全球垃圾邮件和网络钓鱼数据汇总

（APWG）2016年报告显示，超过5500万个网络钓鱼网站，模仿了近700家正式银行或其他金融、电子商务或社交媒体公司等虚假的网站网页，并通过垃圾邮件或其他信息链接吸引诱骗潜在的受害者。[2]面对猖獗的恶意网页钓鱼，亟需运用新技术新手段遏制和阻止。

2、AI+

人工智能用于恶意网页检测主要是通过机器学习的分类和聚类方法。分类方法的逻辑主要是收集数据进行归一化处理后构造分类器识别是否恶意，首先是根据已经标记的URL数据集提取静态特征（主机、URL、网页信息等）和动态特征（浏览器行为，网页跳转关系等），对提取特征进行归一化处理后通过选择决策树、贝叶斯网络、支持向量机、逻辑回归等算法构造分类器来识别恶意网站。聚类方法略有不同，先在网页采集的URL数据集中提取连接关系、URL特征、网页文本信息等特征，再根据聚类算法模型将URL数据集划分为若干聚类，相似度较高的URL数据会在同一聚类中，反之则归为不同聚类，最后对已标记的聚类结果识别待测URL，判断其是否为恶意网页。

3、案例实践

2017年山石网科在Blackhat Asia（亚洲黑帽大会）上发布题为“超越黑名单：运用机器学习检测恶意URL”论文，提出结合域生成算法（DGA）检测的机器学习算法模型，可获得较高的恶意URL检出率。这个模型可对来自全球每周1万条的恶意软件样本库和8万URL地址库数据进行机器学习并检测识别。另外，检测恶意网页还运用了黑名单技术、启发式规则的识别和交互式主机行为等识别方法，如基于黑名单技术的识别方法被应用于Google Safebrowsing，DNSBL，PhishTank等服务，火狐Firefox和IE浏览器则采用基于启发式规则的识别方法。

（六）安全态势感知

1、安全需要

随着互联网对大众生活的渗透纵深影响，网络威胁的演进相应呈现出更隐蔽、波及更广、破坏更强的特点，及时尽早感知发现安全威胁并采取防护措施的难度也在增大。传统安全威胁的感知和分析方法很难适应大数据环境和支持大规模事件分析。因此，加快构建关键信息基础设施安全保障体系，全天候全方位感知网络安全态势，增强网络安全防御能力和威慑能力势在必行。

2、AI+

新型网络安全态势感知的基本原理是利用数据融合、数据挖掘、智能分析和可视化等技术，从宏观角度实时直观显示网络安全状态，并对网络安全趋势做出预测，为网络安全威胁预警和防护提供参考。安全态势感知的关键是对从海量数据中提取要素信息进行预处理和数据融合，再对“提纯”后的信息感知、理解和预测，这与人工智能模式识别与问题预测的任务逻辑具有一致性，因而成就了人工智能技术在网络安全态势感知领域的天然应用。

一般而言，人工智能在安全态势感知系统的运用主要包括五个主要步骤，第一步是提取网络中的安全设备(如防火墙)、网络设备(如路由器、交换机)、服务和应用(数据库、应用程序)的数据，对数据进行标准化和修订，以及标注事件基本特征等；第二步是对在传感器采集环境中获取的数据做去噪、杂质过滤等预处理；第三步融合不同来源的数据；第四步选择人工智能算法进行态势识别、态势理解和态势预测；最后一步完成关联分析和态势分析，形成的态势评估结果常以分析报告和综合网络态势图呈现，用于辅助决策。

3、案例实践

腾讯守护者计划研发的智能金融安全感知系统“灵鲲”力争对非法集资、金融传销等涉众型金融犯罪“打早打小”，识别覆盖率可达90%以上，这套系统基于SAAS服务云的计算能力，汇集大量金融犯罪样本和海量安全云库、网址、应用等动态数据，采用模式识别理论与机器学习方法搭建基础数据层、特征层、决策层等多层级智能决策模型框架，建立起从数据源管理到风险展示的全面态势感知系统；腾讯反传销态势感知系统通过海量监控数据采集、大数据平台储存、自然语言处理分析挖掘、数据接口API对外合法提供数据，全面分析汇总传销态势，该系统目前已协助公安机关破获了善心汇、五行币、云在指尖、星火草原等涉嫌网络传销案

2 数据来源：国际反钓鱼组织发布2016年钓鱼活动报告

件。

安恒信息公司基于其AILPHA大数据网络安全态势感知系统平台为某省打造了农信网络安全态势感知系统平台，通过安全状况视图呈现该省范围内农信资产整体安全情况，系统同时也对机房设备实时监控，及时发现网络入侵、攻击和非法访问。

360企业安全集团基于其EB级网络安全大数据搭建的网络安全态势感知与运营平台实现了针对不同空间、时间、行业和威胁类别的全天候、全方位的网络安全监测、预警和响应，支持10亿级终端和网络安全设备的大规模、自动化应急处置和溯源分析。

网络内容安全篇

（一）舆情监测

1、安全需求

网络信息传播的复杂性、开放性、互动性以及传播环境的隐蔽性使网络平台容易成为舆情危机发酵、扩散的虚拟空间，甚至引发线下社会运动。传统的舆情监测是人工对信息进行筛查、排除，计算机技术的发展使人机结合逐渐应用于信息处理。但现阶段的人机结合呈现出“人机不协调”的问题，使用计算机技术显得过于机械化和浅层次，在处理情感分析和效果检查上并不理想。因此，海量、多维度的舆情信息与舆情监控技术落后之间的鸿沟使舆情信息质量分析能力亟需提高。

2、AI+

人工智能的本质是模拟人脑具有的思维并将其拓展延伸到“拟人思考”和场景模拟中。因此，将人工智能应用于模拟舆论发展过程、识别舆论动态趋势、预测舆论走向成为研究和实践的热点。目前Web挖掘技术、语义识别技术特别是情感分析技术、卷积神经网络（CNN）、支持向量机（SVM）和词频-文档频率（TF-DT）

算法已被应用于网络舆情分析中。[3]

Web挖掘技术主要通过智能信息分析处理技术对内容信息、结构信息和使用信息进行挖掘。该技术主要用于舆情信息采集，可在用户设定的搜索范围内对产生网络舆情的界面和与之相关的界面、内容和使用情况进行采集，提供网络舆情监控分析的数据源。语义识别技术应用于舆情预警领域主要是通过语法预处理、语义内容提取和语义生成三个步骤，将热点、敏感词汇进行聚类和情感分析，自动识别和判断信息的正负属性。自然语言处理中的词频逆文档频率技术（TF-IDF）是一种用于信息检索与数据挖掘的常用加权技术，该技术用于文本特征提取、比较文本相似度、文本分类。该技术的最大优势在于算法简单，且对文章所有元素进行综合考量，信息分类处理的速度快，可缩减危机舆情发现和响应之间的空窗期。通过上述主要的人工智能技术，舆情监控的即时性、有效性得到显著增强。

3、案例实践

腾讯安全团队联合中国电科开发了一款智能化的“网络社会安全风险态势系统”，通过针对网络社会安全构建相应评价指标体系，基于系统提供的数据分析和智能预测，对实现对社会及政府治理中重点关注的公共安全问题进行舆情跟进和及时预警。

中科点击公司研发的军犬舆情监测系统通过强大的信息采集系统对数据进行挖掘，能对多国语言的网络信息舆情进行监测。军犬舆情能够智能提取网页中的有效舆情信息，对具有关联性的多个网页内容进行自动合并、自动提取，从而达到全方位舆情预警的功能。

（二）网络谣言治理

1、安全需求

网络谣言具有传播范围广、传递速度快、社会影响较大的特点。传统的网络谣言治理模式主要包括法律法规规制、公民自主约束上网行为、提高网民媒介素养、举报疑似谣言、网络平台辟谣等，尚存在发现难、举证难、认定难、易反复等困

3 魏晓光, 孙康琛, 张涛,等. 基于人工智能的网络舆情监管模式创新探究[J]. 产业与科技论坛, 2017, 16(4):48-49.

难。在当前网络信息量和移动端用户量剧增的情况下，谣言治理模式亟需升级，在谣言发现、谣言鉴别、辟谣信息推送等方面需提高响应效率，扩大辟谣科普信息影响力，以应对当前瞬息万变的信息流散。

2、AI+

依托互联网全网优质数据资源，通过自然语言处理技术识别文本、图像和视频中的需筛查内容，用AI深度学习技术不断完善算法模型，人工智能为治理网络谣言提供了新的技术解决路径。人工智能治理网络谣言的主要步骤为鉴定、分类运算、人机结合发布辟谣信息和定向推送读者，以遏止谣言扩散。首先，对谣言内容属性和传播特征进行提取，内容属性包括对谣言的文本关键词、发布时间、发布地点进行甄别比对，传播特性是指一般的谣言传播是由点到面进行核裂变式传播，与一般的新闻传播方式不同。然后，通过分析谣言特性建立算法模型，获得事件相关的谣言与非谣言数据训练集，再运用分类算法对测试集进行分析，之后采取人机结合通过权威机构或平台发布专业辟谣信息。此外，还可利用算法推荐定向将辟谣信息推送给谣言易感读者，遏制谣言二次传播。

3、案例实践

腾讯公司基于其海量的用户量以及人工智能技术领域的经验积淀打造了全套辟谣产品矩阵。该矩阵包括腾讯新闻较真平台、微信公众平台辟谣中心、微信安全中心、腾讯内容开放平台企鹅号辟谣机制和手机管家安全头条等。这几大矩阵产品通过AI技术打造优质的辟谣数据库，实现智能查询甄别谣言，借助机器算法触达谣言易感人群，基于阅读或投诉谣言的类型标签进行精准推送辟谣防谣。据《2017腾讯公司谣言治理报告》数据显示，腾讯辟谣产品矩阵在2017年总辟谣阅读次数达8亿次，拦截谣言超5亿次，累计发布辟谣榜单及研究报告76个，辟谣合作专家和机构达1304家。

此外，作为全球最大社交平台Facebook也很早就将人工智能用于虚假信息治理，2015年Facebook推出一套智能系统，借助大量用户的群体标记，对已标记虚假信息链接进行降级处理，优化机器算法后更大大降低了人工甄别虚假新闻的比例，提高假新闻监测效率和拦截准确率。

（三）不良信息内容检测

1、安全需求

随着移动互联网的普及与技术发展，网络信息生产更加便利，互连网络空间中的信息总量十分庞大，视频化的趋势日益凸显，网络信息内容的产生与传播具有实时、海量、多态、流动等特性，内容生产与传播的同步性导致对网络内容的管理往往很难预测和前置，给网络信息内容治理带来极大的挑战。传统的网络内容治理工作，在音频、视频等媒体信息处理中存在发现难的问题，且内容治理依赖人工审核做确认，投入成本高、效率低，也无法适应新形势下信息量巨大的现状，内容治理模式亟需革新。

2、AI+

信息内容形式有文字、图像、视频和语音，人工智能技术已能够在应用于上述不同形态的网络信息内容治理中，大大提高信息内容治理的效率和有害信息内容的覆盖。

主要通过基于神经网络的CNN、RNN、DCNN、NMT和特征金字塔网络等深度学习技术，可在文本内容检测、文本分类、图像识别、OCR、人脸检测、视频识别、语音识别及关键词唤醒等方面发挥重要作用，用以检测和识别色情、辱骂、枪支、赌博、暴力、恐怖主义等违法犯罪信息。

以图像/视频鉴黄为例，我们通过深度学习，将标注过的海量的正常、性感、招嫖、色情样本预处理后输入“机器”进行学习，并持续进行样本喂养，迭代调整训练效果，待训练过程收敛后输出识别模型，该模型应用在各大场景检测色情内容，可大大提高色情信息识别的准确率以及识别效率，有效减少人工审核的投入，提高内容治理的效率和效果。与此同时，我们还会通过OCR的方式提取图片和视频中的文字，并进行NLP的语义理解识别招嫖和色情；尤其值得一提的是，对于户外拍摄、直播等短视频没有明显漏点和特别模糊的情况下，启用了提取视频的语音，做色情语音的识别，有效地提升了色情内容的覆盖。

3、案例实践

在不良信息检测方面，腾讯、微博、阿里巴巴、百度、Facebook、Google等各大互联网公司都在积极开发提供“AI+”的不良信息检测服务。

腾讯公司的冰鉴引擎已经具备了基于AI完善的端对端多种能力，包括色情/招嫖识别、图片语义相似度、暴恐识别、OCR、音转文/关键词唤醒等，已经全面应用于腾讯业务场景中。其中，冰鉴引擎对于图片/视频的色情识别准确率达99.9%以上，覆盖85%左右，对短视频和直播平台上的不良信息进行高效检测；对于图片/视频的暴恐识别方面，更是引入先进的Image Caption技术，很好地应用到了具体业务中；OCR方面，不仅能够识别身份证、印刷文字、营业执照等标准文件类型，而且对于UGC产生的对抗混排、大小字体不一，不同字体图片有很好的效果，特别是首创通过类似语音关键词唤醒的技术极其快速地检测图片是否包含色情等不良信息中特定的关键词；通过CNN语义相似度能够实时快速地处理一些突发事件，譬如蓝鲸游戏期间，我们通过迁移学习的技术实时发现了大量有害群体。语音领域，通过关键词唤醒技术能够快速实时检测语音中辱骂的一些词语，并做响应的惩罚；同时，为加强民族间的交流，研发了支持藏、蒙、维、英与汉语同传的语音技术。

（四）诈骗信息打击

1、安全需求

骚扰诈骗电话和垃圾短信一直是困扰手机用户的难题，其中潜藏的诈骗风险让用户饱受财产损失之苦。据腾讯手机管家与腾讯安全联合实验室移动安全实验室发布的《2018年上半年手机安全报告》显示，2018年上半年，腾讯手机管家共收到用户举报骚扰电话近1.43亿次，其中诈骗电话近2970.52万。诈骗信息不仅对个人财产造成损失，也对社会稳定造成影响。因此，对诈骗信息必须借助更深、更广、更有效的打击方式。

2、AI+

人工智能技术运用于打击诈骗信息和欺诈行为主要是通过数据采集、数据分析和决策引擎实现。数据采集是在遵守国家法律法规和监管要求下，通过设备指纹、网络爬虫、生物识别、位置识别、活体检测等技术从PC端或移动客户端获取用户相关必要数据。之后，结合数据分析建立反欺诈经验规则库不断为机器学习供给数据“饲料”，利用有监督学习、无监督学习和半监督机器学习模式对数据进行训练分析，构建适合模型预测欺诈行为。决策引擎是数字反欺诈的核心，主要是通过决策引擎将信誉库、专家规则和反欺诈模型等各类反欺诈方法有效整合。比如，CEP引擎（复杂事件处理系统）就是实时计算分析（过滤、关联、聚合）与欺诈案件相关的多类事件之间的关联性，精确分析用户意图，还原事件场景，降低误杀率。

3、案例实践

腾讯公司基于其黑产对抗经验及技术和反欺诈AI模型建造能力，打造“宾果反诈骗防控系统”，该系统通过海量学习警情和大数据分析能力，自主提取警情中作案手法、通信行为、网络特征、资金流向等特点规律，实现智能建模、智能运算、智能预警，在诈骗事前、事中、事后等环节起到预警、分析作用。通过对警情、通信、网络、金融等领域大数据的深度学习，宾果系统可自动发现电信网络诈骗犯罪行为并自动预警，自2018年3月21日上线以来（截止9月上旬），宾果反诈系统预警已超过50000起，准确率超过99%，覆盖全国31个省区市，累计为民众避免损失20亿元。其次，宾果系统可实现对电信网络诈骗窝点、人群的智能聚类，为警方开展刑事打击提供线索参考。宾果反诈骗防控系统投入使用以来，全国重点诈骗区域诈骗号码大幅下降，降幅接近70%，有效遏制了诈骗态势蔓延。此外，综合腾讯多年安全能力，宾果系统还能够对已发案件进行智能分类、特征刻画、人员扩展和团伙聚类，通过系统机器学习和自动运算，排查诈骗团伙窝点位置、规模、人员、作案手法等，为警方开展针对性刑事打击提供有力参考依据。

此外，2017年面对电信诈骗中发现难、抓获难、定罪难的伪基站，蚂蚁金服集团与公安部刑侦局合作开发了“伪基站实时监控平台”，基于大数据支持和生物识别技术在全国实施实时监控，能做到精确到50米内的定位。目前，该平台已覆盖全国31个省（直辖市、自治区），一年内协助公安机关打击了37个伪基站涉案团伙。[4]

4 综合整理自网络关于2017年广东警方“飓风1号”专案行动相关新闻报道。

（五）金融风险控制

1、安全需求

金融风险预警是稳定金融市场发展的基础性环节，传统的金融预警基本上是通过人工搜集相关资料，凭经验分析风险因素变动和预估风险偏颇状态，但面对当前金融业务和客户量的激增，数据异构性所带来的问题也不断显现，基于专家经验的传统金融风控方式显得力不从心。

2、AI+

深度学习生成的框架模型对于处理海量异构数据有着优秀的表现，特别是在提取文本、图像、视频等大量非结构化数据中的特征方面，优化数据质量，神经网络、专家系统、支持向量机、决策树等算法在提升金融风险控制决策效率和准确率方面卓有成效。基于规则语言做出决策的专家系统一般多用于金融机构对企业信用贷款授信申请做出决策，减小贷款风险。支持向量机算法则常被应用在个人、企业信用数据评估，提高信用风险分类精度。拟合逻辑决策树模型在信用级评审中的应用可将大量经典案例与经验逻辑相结合，使风控规则模型更趋于真实，满足灵活场景下的风控辅助决策，大大节约模型优化的时间和坏账成本。

3、案例实践

支付宝借助大数据和AI技术打造了AlphaRisk智能风控引擎。该引擎主要由Perception（风险感知）、AI Detect（风险识别）、Evolution（智能进化）、AutoPilot（自动驾驶）4大模块组成，结合了人类直觉AI(Analyst Intuition)和机器智能AI(Artificial Intelligence)运行AutoPilot(自动驾驶)。AutoPilot利用智能算法推荐最优核身策略进行精准推送，降低因盗取身份而带来的金融风险。例如若AlphaRisk识别到支付宝账户存在手机丢失风险，AutoPilot能够自动升级核身方式用人脸或指纹智能方式校验。另外，AlphaRisk智能引擎能够对每个用户的每笔支付进行7×24小时的实时风险扫描；通过不断新增的风险特征挖掘和优化算法迭代的模型，自动贴合用户行为特征进行实时风险对抗，准确识别用户的账户异常行为，完成风险预警、检测、管控等流程仅不足0.1秒。

安恒信息公司开发的涉众型金融风险监测预警处置平台，通过机器学习对数据进行挖掘和建模分析，再结合专家研判服务将金融信息划归为红、橙、黄三类等级纳入动态监测系统，根据检测结果形成预警。此外，美国华尔街交易所和伦敦证券交易所也布设了基于机器学习算法建立的智能监管系统，用来检测和识别交易风险行为。

物理网络系统安全篇

（一）智能电网建设

1、安全需求

国家电网信息系统一直是国家关键信息基础设施之一，关系国计民生。随着互联网泛在发展，电力信息系统更容易遭到攻击，一旦电力信息系统遭到侵害，将会导致整个国家电力系统和其他使用电能的关键设施陷入瘫痪，其破坏力不亚于对国家实施武力袭击。因此，建设安全、智能、高效的电网系统，保障海量、多类型、复杂度高的电网数据安全，维护电网系统的运转安全已经成为国家基础设施建设和保障要务。

2、AI+

发电侧和电网侧的安全保障是电网维护工作的重中之重。借助人工智能技术，在发电侧对发电设备进行故障检测、诊断与预警，结合储能AI技术可以接入清洁能源并协调优化其运行，实现清洁能源并网的“即插即用”式仿真分析。在电网侧，基于数字仿真的大电网人工智能分析，可实现电网状态感知、输变电设备巡检与诊断、基于微扰动的电网动态进行特征提取、扰动源识别等。目前，人工智能已经应用于智能调度大数据应用架构、广域优化调度、负荷预测、全周期设备管理等场景，还对优化电力生产、输送、调度、分配以及消费发挥着重要作用。

3、案例实践

法国的阿尔斯通（ALSTOM）公司提出将利用大数据分析技术对实时数据、电表数据、财务数据等海量且重要的数据进行整合分析，支持电网运行和用户服务。美国C3公司运用大数据系统每小时处理50亿条数据，这意味着每台电表平均可产生300美元的经济效益。[5]中国智能电网建设中使用Agent技术及多Agent技术等分布式人工智能技术可准确并稳定地进行电价预测，电力系统继电保护方面取得显著呈现，智能电网建设已大力助力于国家电网基础设施保护。

（二）智能交通调控

1、安全需求

随着城市化的不断深入，中国正在快速步入汽车社会，机动车辆保有数量迅速增长，使城市尤其是大型城市面临着严峻的交通压力。交通的拥挤不仅会造成安全事件的频发，也会带来经济损失，严重时将导致城市功能的瘫痪。传统的城市交通调控主要以人工检测、调配为主，但面对当前的压力仅凭人工调度已经不足以运转庞大的城市交通体系。因此，运用智能技术调控城市交通正逢其时。

2、AI+

随着信息技术的进步，智能化交通系统的功能、效用不断优化，当前人工智能带来的技术突破已对该系统进行了全面升级。基于历史数据法、时间序列法和神经网络仿真法已能完成对交通流量进行预测，使得人、车、路之间的联系得以加强。驾驶人员可实时、快速的掌握一手路况、交通流量信息，有效地缓解了大城市的交通拥堵状况，减少交通事件发生的频次。同时，通过多尺度融合、多种平衡采样策略、高精度优化检测框已经大大提升了车辆检测技术。

3、案例实践

在人工智能运用于车辆检测上，腾讯图优设计的车辆属性识别产品已能实现对22种车型、300多种品牌、4000多种车系年款、11种车身颜色、以及年检标、遮

5 闫湖, 狄方春, 袁荣昌,等. 电网智能调度中的大数据及应用场景研究[J]. 电力信息与通信技术, 2014, 12(10):7-12.

光板等特征区域进行高达90%的识别率，支持新能源车等各种种类的车牌识别率达99%。对车辆的精准识别大大提高了抓捕交通肇事者的效率和准确率。另外，腾讯科恩实验室致力于保障智能网联汽车“大安全”，在智能车辆系统检测领域已取得重大突破，也在同步开展自动驾驶相关的信息安全研究。近年来曾通过移动网络识别、准备和实施攻击场景等技术成功发现高级智能车辆特斯拉的重要漏洞，并遵循负责任的漏洞披露方式把发现的漏洞威胁情况提交给特斯拉官方，供其完善智能防御系统，保障智能车辆安全。在当前智能攻防竞争态势中，这种先于黑客发现漏洞的自主安全保障能力是极其需要的。此外，百度公司提出的智能交通信号灯设计，也体现着人工智能在优化交通系统中的重要作用，通过机器学习、人脸识别和语音播报等技术让人们出行更加方便。通过人工智能技术建立的道路交通管制、公共交通指挥与调度、高速公路管理和紧急事件管理等应用，将助力国家交通系统的安全保障。

（三）城市安防监控

1、安全需求

城市安全是城市发展的重要基石，视频监控则是保障城市安全的重要手段。传统的视频监控系统，只能记录图像（含声音），对于个体、人群的移动轨迹需要人工来查看、刻画，程序复杂且工作量大。面对当前城市数量和城市人口的激增，传统视频监控已经难以完成海量的视频收集和分析处理。

2、AI+

基于人工智能技术，人体动线的自动检测跟踪已成为现实，它以人头、人体融合跟踪的模式来提高跟踪准确率，且加入多种算法和自适应阈值来降低误差。同时智能视频系统可以将平时的场景、模式、处理方式等数据保存至智能化监控系统平台中，进行数据喂养，随着运行时间的增加系统经验值也将不断上升，其智能化程度不断加强。

3、案例实践

在目前应用中，城市监控系统已经给家庭、校园和大型公共场合的安全提供了

重要保障。安装家庭安防系统，让家庭保持24小时被记录，可使用人脸识别和模糊识别对闯入家中的不法分子实时预警并第一时间报警。校园里安装智能视频监控，实现对周界防范预警功能，通过计算机可视化对校园内和校园周边的暴力事件或潜在暴力事件进行识别和防范。在大型公共场合，智能视频能够在人群聚集区域有效获取实时画面监控和人流量反馈，防止大型公共安全事件的发生。另外通过跨省、跨地域的智能监控视频能够对犯罪分子进行查找，对被拐卖的儿童进行寻回。腾讯优图实验室研发的跨年龄人脸识别技术被用于“QQ全城助力”的公益项目，已协助警方和公益组织寻回302个被拐和走失儿童。

第五章 展望

（一）人工智能安全将成为产业发展最大蓝海

可以预见的是，智慧城市、工业互联网、自动驾驶等将在未来十年全面普及，安全将成为各类智能创新应用最核心的痛点需求，也是人工智能技术最重要的应用领域。为此，各国都会全面加大人工智能在安全领域的应用，如智慧城市建设中的安防领域，其产出的海量数据和其防护逻辑与AI技术逻辑的高度自洽，使AI技术能够天然应用于安全。

（二）人工智能本体安全决定安全应用进程

人工智能在助力解决各领域安全问题的同时，其自身的安全性也越来越重要。如何保障合理的运用AI技术一直是人类面临的难题。算法、数据、物理载体的安全性决定着AI的本体安全，是AI助力网络空间安全的基本前提。因此，AI技术助力安全的发展如同DNA的两条单链，一条单链是在人工智能技术不断与更多产业融合下，人工智能保障安全需求迅速增长，第二条单链则为对保障人工智能本体安全的需求增长。两条单链相互催生，构成了人工智能技术助力安全的螺旋式发展。

（三）“人工”+“智能”将长期主导安全实践

长期来看，人工智能技术只是辅助而非替代人类的关键判断，其中安全决策尤为复杂，更是人工智能无法完全替代的领域，因此人类决策与机器智能将长期并存。例如，Facebook对网络新闻的真假判断仍然是在机器学习对信息进行降级处理后进行人工审查作出最终判断。在网络谣言治理领域，人工智能应用依然主要采用机器+专家审核模式。因此，基于人工智能对于复杂的社会关系、情感识别、价值判断的能力依然不足，智能和人工结合的人工智能模式将长期主导应用实践。

（四）人工智能技术路线丰富将改善安全困境

机器学习、深度学习等人工智能技术高度依赖海量数据的“喂养”，但是，数据采集与隐私保护之间已经形成囚徒困境。因此，随着人工智能技术路线的不断丰富和发展，可以根据用户需求和应用场景，有针对性的选择人工智能技术，可以避

免智能应用对数据资源的过度依赖，更好地保障网络空间安全。

（五）网络空间安全将驱动人工智能国际合作

面对共同的威胁是国际合作的重要前提。人工智能时代，无论是在应对网络犯罪、网络攻击等安全威胁，或是无人驾驶、智慧城市的安全保障，都需要各国技术标准、信息资源、应对机制等方面的匹配协同。其中，面对技术规范、行业标准、法律法规尚未健全的人工智能新领域，制定统一的安全监测标准、安全防范架构、安全评估体系需要各国参与共同协调，网络空间安全成为人工智能国际间合作的重要领域。