

AI 安全的演变：从风险到威胁

薛峰 韦韬
蚂蚁集团

关键词：AI 安全 威胁对抗

引言

近几年，无论是网络安全还是人工智能（AI）都呈现蓬勃发展之态势，并且都上升到了国家战略的高度。2018年4月，习近平总书记在全国网络安全和信息化工作会议上提到，网络安全牵一发而动全身，深刻影响政治、经济、文化、社会、军事等领域的安全，没有网络安全就没有国家安全。同年10月，习近平总书记更是强调人工智能是新一轮科技革命和产业变革的重要驱动力量，加快发展新一代人工智能是事关我国能否抓住新一轮科技革命和产业变革机遇的战略问题。在双轮战略的驱动下，这两者的交叉领域的研究也正在如火如荼地开展，并产生了很多“化学反应”，从目前来看主要有三个方向，分别是人工智能自身安全、人工智能助力安全、人工智能衍生安全。人工智能自身安全，顾名思义就是指AI本身的风险问题，在某些语境下可以简称为“AI安全”，类似于系统安全、网络安全的概念。例如经典的大熊猫对抗样本攻击案例，恶意攻击者只需要修改大熊猫图片上的几个像素点就能导致人工智能系统识别错误，本应该识别出大熊猫却错误地识别成了长臂猿，而且置信度很高。人工智能助力安全即用AI的方法来解决安全的问题。这个领域起步更早，在本世纪初就有科学家基于人工智能技术做垃圾邮件的检测，取得了非

常好的效果。近几年在业务安全领域更是涌现出非常有意思的案例，如智能风控、智能反洗钱、智能垃圾内容识别等，这些技术极大地提升了风险识别的准确率和召回率。系统网络安全领域也不例外，衍生出了AISecOps智能安全运营的新理念。人工智能衍生安全是指AI技术导致某些其他领域不安全了，比如最近流行的AI换脸技术，它能够把照片或者视频中的人脸换成另外一个人，而且比Photoshop更加方便，这类技术使得内容安全领域面临前所未有的挑战。

人工智能的安全风险

任何一项技术都伴随着安全风险，人工智能技术也不例外，近年来，越来越多的安全专家和人工智能专家揭示了大量AI技术自身的安全问题。安全专家称之为“AI安全”。保障系统安全运行是安全专家的职责，而揭示风险是实施风险控制闭环的关键步骤，

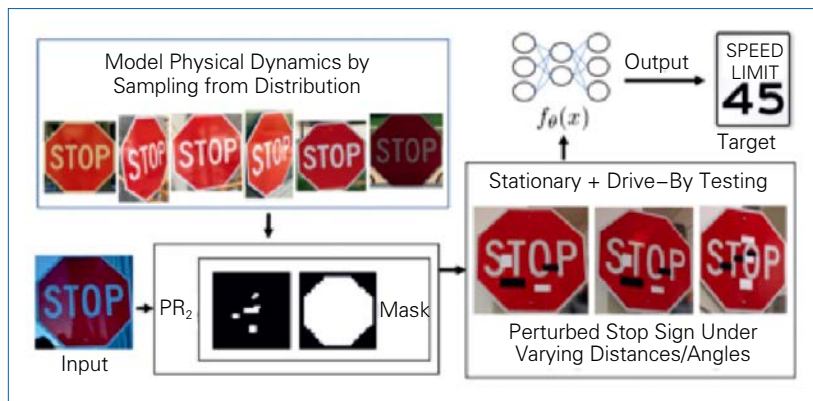


图1 交通标志被干扰示例

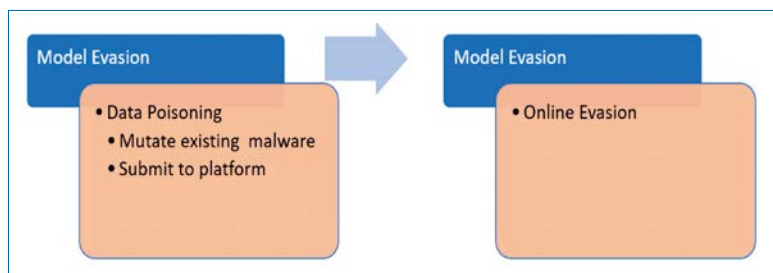


图2 逃逸攻击案例

于是就出现了各种“花式吊打”AI的方法。图1是一个交通标志的例子，原本应该被AI识别出“Stop”的停止标志，但被加上各种扰动之后，最后AI识别出来的结果却是“限速45”，可以想象，如果这种不安全的算法出现在自动驾驶汽车上会带来多大的灾难。

安全领域有一个著名的CVE(Common Vulnerabilities & Exposures)漏洞数据库，它由非营利性组织MITRE运营。2020年10月，MITRE发起了Github开源项目“advmlthreatmatrix”，该项目旨在系统化地梳理针对AI的威胁矩阵，目前已经有蚂蚁集团、微软、IBM、卡耐基梅隆大学等十几家机构和高校参与。图2是其中的一个攻击案例，攻击者首先下载一个流行的恶意软件并埋入精心构建的后门，该操作被称为“数据毒化”，然后攻击者将这个被毒化的恶意软件再次上传回平台，并宣称是新的恶意软件样本，而一旦该样本被人用来训练恶意软件识别模型，则该模型也就被植入后门，最终效果是凡是带有特殊信息的恶意软件都不会被该模型识别了，即实现了“逃逸攻击”。

这类问题在人工智能的专业领域里被称之为鲁棒性问题、泛化性问题，AI系统的正确率往往不是100%，即使在某些任务下(如人脸识别)正确率已经超越了人类，但是AI系统太容易出错了，尤其是在“对抗环境”下，有来自黑灰产的恶意攻击。随着无人驾驶、机器翻译等AI技术逐步民用，如果安全性要求不达标，势必会影响AI技术的普及和发展。

2020年国内有两大AI安全白皮书问世，分别是由中国信息通信研究院牵头发布的《人工智能安全框架(2020年)》和由浙江大学-蚂蚁集团金融科技研究中心牵头发布的《人工智能安全白皮书(2020)》。

前者以实用性、可用性、整体性、前瞻性为设计原则，提出了人工智能安全框架，包括安全目标、安全能力、安全技术、安全管理，为提升人工智能安全提供了有益指引。后者侧重于从学术界、工业界的共同视角来阐述人工智能安全，系统性地归纳和总结了AI模型、AI数据与AI承载系统面

临的风险和加固手段。

AI承载系统这一层属于网络安全领域范畴，如TensorFlow的各类漏洞。AI数据层和AI模型层则涌现出新型的风险，以下介绍三种典型场景。

数据投毒攻击：指攻击者通过在模型的训练集中加入少量精心构造的毒化数据，使模型在测试阶段无法正常使用(可用性问题)，或协助攻击者在没有破坏模型准确率的情况下入侵模型(完整性问题)。毒化攻击可以攻击几乎所有算法，包括计算机视觉域算法、自然语言处理域算法、语音域算法、联邦机器学习、推荐、搜索等。此外还会衍生出后门攻击，以图像分类为例，攻击者通过精心构造带触发器的图像数据集毒化模型，使得模型分类错误，将图片分类到攻击者指定的类别。

对抗样本攻击：指利用对抗样本对模型进行欺骗的恶意行为。对抗样本是指在数据集中通过故意添加细微的干扰所形成的恶意输入样本，在不引起人们注意的情况下，可以轻易导致机器学习模型输出错误预测，例如上文提到的交通标志的错误识别会造成无人驾驶汽车做出错误决策从而引发安全事故。对抗样本的发现严重阻碍了AI技术的广泛应用与发展，尤其是对于安全要求严格的领域。因此，近年来对抗样本攻防技术吸引了越来越多的目光，成为研究的一大热点，涌现出了大量的学术研究成果。

模型窃取：模型窃取攻击是一类数据窃取攻击，攻击者通过向黑盒模型进行查询获取相应结果，窃取黑盒模型的参数或者对应功能。被窃取的模型往往是拥有者花费大量的金钱和时间构建而成的，对拥有者来说具有巨大的商业价值。模型的信息一旦遭到泄露，攻击者就能逃避付费或者开辟第三方服务，从而




	Selection-based captchas	Slide-based captchas	Click-based captchas
Examples			
Providers	facebook.com, 12306.com, google.com	geetest.com, tencent.com, 163.com	geetest.com, tencent.com, 163.com
Attacks	Sivakorn <i>et al.</i> [44], Ya <i>et al.</i> [49], This paper	This paper	This paper
APIs	GoogleAPI, TencentAPI, AliAPI, MirosoftAPI	—	BaiduOCR, GoogleOCR, TencentOCR, AliOCR, Face++OCR
Captcha-solving Services	ruokuai, yundama, hyocr, 2captcha, AntiCaptcha, Decaptcha, imagetypers	ruokuai, hyocr, dama2	ruokuai, yundama, dama2

图3 学术界总结的对各种类型验证码的破解方法

获取商业利益，使模型拥有者的权益受到损害。攻击者如果成功窃取模型，就可以进一步部署白盒对抗攻击来欺骗在线模型，这时模型的泄露会大大提高攻击的成功率，造成严重的安全问题。

更详细的内容读者可以参阅发布在 Github 上的白皮书《人工智能安全白皮书（2020）》，我们坚信随着国内的 AI 安全大赛的火热，以及各大高校和科研机构不断的加码投入，相信在未来五年内，AI 自身的安全性一定会得到飞速发展，AI 技术也将在各行各业持续创造可靠的生产力。

基于 AI 的威胁对抗

虽然 AI 自身风险类型较多，但是在实际情况中真正形成威胁态的还相对较少。威胁是已经发生或者即将发生的安全事件，是有实际对手的，如黑灰产、黑客、不法资本、恶意竞争对手，甚至出于政治目的的国家级对手。相对于 AI 安全域，网络系统安全域、数据安全域、业务安全域正面临着前所未有的挑战。

安全的本质是对抗，威胁对抗的核心三要素是对手、资产、对抗体系，接下来我们从这三个视角系统化地介绍安全与 AI 的交叉点，从而更好地理解 AI 衍生安全以及 AI 如何助力安全。

威胁对抗的第一要素“对手”指的就是通常意义上的攻击者。在军事战争中如果连敌人都不了解，注定会走向失败，安全领域同样如此，一定要擅于对

“对手”进行威胁建模。特别是在甲方安全中，如果过度强化对手的能力会导致无意义的成本消耗，而如果过度低估对手的能力则会导致一系列安全事件。安全工作者大部分时间对抗的都是黑灰产，而不法资本、恶意竞争对手、国家级攻击等是非常稀疏的。黑灰产一般分为两派，一派是有组织的黑产，他们会不断开发新的黑产工具、平台、基础设施，甚至完整的下游洗钱链路；一派是低端黑产，低龄化趋势明显，他们会利用黑产工具不断地进行非法获利。当下一个很重要的趋势就是黑灰产正在拥抱 AI，也就是人工智能衍生安全，从黑产的角度讲就是所谓的 Weaponize AI（人工智能武器化），他们背后的诉求就是尽可能把攻击行为自动化，以提高攻击效率从而谋求更大的利润。

除了 AI 换脸技术之外，最典型的案例就是用 AI 技术破解验证码了。图 3 是学术界总结的对各种类型验证码的破解方法。验证码本来是用来区分机器人和人类的，如文字验证码、图形验证码等，而随着 AI 技术的发展，机器已经能够读懂文字、图形，与人之间的差距越来越小，甚至在某些领域的能力已经超越了人类，所以目前用 AI 技术破解验证码大行其道，这些海量的机器行为被用来进行各种“薅羊毛”、账密撞库¹等攻击。

威胁对抗的第二要素“资产”是指安全工作者保驾护航的对象。对一个大型互联网公司来说，资产形态很多，包括系统、网络、硬件等基础设施以及

Web 应用等。企业安全中有安全域的概念，如基础设施安全、应用安全、数据 & 隐私安全、业务安全、办公网安全等，每个安全域都有相应的安全工作者负责。安全工作者的日常工作中很大一部分就是对资产进行风险识别，如果发现潜在问题则进行管控，甚至推动复合治理，整体目标就是不断降低暴露在外的漏洞、风险面，不断逼近攻击者无漏洞可攻击的状态。“资产”这一层与 AI 的结合点包括 AI 自身风险，也包括 AI 助力安全，如基于强化机器学习挖掘基础设施中的代码漏洞、基于 NLP 技术/CV 技术识别敏感的数据资产，本质上都是加快风险发现的速度进而做好风险管控。

威胁对抗的最后一个要素是“对抗体系”。世界上没有一个绝对安全的系统，快速消除威胁的方法就是对抗，对抗作战体系对于企业安全极其重要。2014 年是互联网金融元年，数字世界的高价值资产吸引了大量黑灰产进行“薅羊毛”、盗账户、盗卡等违法犯罪活动，业务安全备受挑战。但挑战就是机遇，也就是从那几年开始，智能化的风控、智能化的反洗钱开始兴起，到现在发展得已经相对成熟。AI 技术带来的增益可以归纳为两点，第一是针对已知的攻击手法，AI 技术能够提升识别的精度，规则体系下的阈值往往是最难设置的，而 AI 技术可以从已知的黑样本中学习出恰当的模型，从而提升识别精度；第二个增益点是 AI 技术能够稽核出来一些未知的攻击手法，这得益于聚类、异常检测等数据挖掘技术，成功化被动安全为主动安全。

近几年，基础设施安全领域也开始逐步拥抱 AI，很多安全产品都带上了“智能”“大脑”等字眼，如智能反入侵、安全大脑、用户和实体行为分析 (User and Entity Behavior Analytics, UEBA)、智能反爬虫等。2020 年，奇安信集团、清华大学、蚂蚁集团联合举办了 DataCon 大数据安全分析竞赛，旨在利用人工智能等新技术方法对不同场景下的安全问题进行智能分

析，包括 DNS 恶意域名分析、恶意代码分析、僵尸网络分析，等等。

当然，新技术在未来五年全面落地还需要克服诸多挑战，第一大挑战就是数据问题，基础设施安全领域的的数据从 3V (Volume-Variety-Velocity) 模型来看比业务安全领域复杂得多。在数据量级 (volume) 维度上，业务安全领域的的数据量与数据库的条数成正比，一般十亿级就已经很大了，而基础设施安全领域的的数据量与网络流量、主机日志等成正比，一般是万亿级，这对大数据计算引擎、数据科学计算引擎提出了相当高的要求。在数据类型 (variety) 维度上，业务安全领域由数据库存储，属于结构化的数据，而基础设施安全领域非常复杂，有半结构化的日志文本数据，也有需要做分类分级的图像数据，甚至有些资产还没有被管理起来，需要去部署监控智能体 (agent)，而智能体的数据采集率、送达率、时效性等都具有挑战。第二大挑战是 AI 技术的鲁棒性问题。对于模型自身的安全性，当用一个存在漏洞的模型去对抗黑灰产时，容易被黑灰产躲避过去。如在钓鱼邮件场景中，如果识别模型不够鲁棒，那么将无法检测出黑灰产精心炮制的钓鱼邮件。因此对抗机器学习技术方向仍然有很大的发展空间。第三大挑战就是 AI 技术的“推理”能力。AI 技术虽然能够检测出未知攻击，但是目前还需要安全专家判断其攻击行为的真假。当下的 AI 技术无论是深度学习，还是经典的机器学习，本质上是做模式识别 (pattern recognition)，特别擅长寻找数据之间的关联，而如何把知识融入进去，让 AI 能够实现自动化推理，还需要较长时间的技术积累。所以在未来五年甚至十年之内，很难达到 AI 完全替代人的理想状态。而安全专家把 AI 作为工具来提升工作效率的趋势将持续存在，这正如军事领域的发展，从冷兵器时代的肉搏到工业时代开始有机器代替人力，再到信息时代有计算机算力的加持，虽然装备越来越先进，但是部队、总政、总参、特种兵等依

¹ 撞库是黑客通过收集互联网已泄露的用户和密码信息，生成对应的字典表，尝试批量登陆其他网站后，得到一系列可以登录的用户。很多用户在不同网站使用的是相同的账号密码，因此黑客可以通过获取用户在 A 网站的账户尝试登录 B 网站，这就可以理解为撞库攻击。

然存在。也正是有了这些新科技的进步，单兵作战的能力和几百年前不可同日而语。

结语

安全对于AI技术来说是一个非常好的练兵场景，各种问题和挑战将会促进AI技术向前发展，走向强人工智能时代。同样，AI技术对于安全来说也是非常核心的生产力，在对手都在拥抱AI的趋势下，安全工作者没有理由不去拥抱AI以提升单兵作战的能力。没有网络安全就没有国家安全，虽任重道远，但未来光明。相信AI安全这个交叉领域未来五年会蓬勃发展，为“十四五”规划的全面胜利贡献力量。 ■



薛 峰

CCF 专业会员。蚂蚁集团对抗智能负责人。主要研究方向包括 AI 安全、隐私计算。
gknlfexxx@gmail.com



韦 韬

CCF 专业会员。蚂蚁集团副总裁。主要研究方向为系统安全、数据安全、隐私保护与隐私计算、AI 安全、企业安全。
lenx.wei@antgroup.com