CSA GCR cloud security alliance ®
GREATER CHINA REGION

AI安全白皮书

联盟简介

云安全联盟（Cloud Security Alliance,CSA）是中立、权威的全球性非营利产业组织，于2009年正式成立，致力于定义和提高业界对云计算和下一代数字技术安全最佳实践的认识，推动数字安全产业全面发展。

云安全联盟大中华区（Cloud Security Alliance Greater China Region, CSA GCR）作为CSA全球四大区之一，2016年在香港独立注册，于2021年在中国登记注册，是网络安全领域首家在中国境内注册备案的国际NGO，旨在立足中国，连接全球，推动大中华区数字安全技术标准与产业的发展及国际合作。

我们的工作

联盟会刊下载地址
了解联盟更多信息

加入我们

CSA大中华区官网（https://c-csa.cn）

点击会员

加入联盟

填写相关申请信息

成为CSA会员

JOIN US

致谢

《AI 安全白皮书》由 CSA 大中华区 AI 安全工作组专家撰写，感谢以下专家的贡献：

工作组联席组长:

黄磊　　王维强

主要贡献者：

黄磊	王维强	何伊圣	卞超轶	李岩	陶瑞岩
刘广坤	郑国祥	张淼	陈洋	郭建领	唐科伟
崔鉴	段阳阳	冯明	邹旭	郝伟	黄国忠
林兵	邢海韬	杨浩淼	杨喜龙	张亮	孟昌华

审校组:

黄磊	王维强	黄连金	刘广坤	何伊圣	张亮
李岩					

研究协调员：

黄家栋

贡献单位:

中国电信股份有限公司研究院	蚂蚁集团安全实验室
北京百度网讯科技有限公司	杭州安恒信息技术股份有限公司
北京启明星辰信息安全技术有限公司	华为技术有限公司
联通（广东）产业互联网有限公司	北京沃东天骏信息技术有限公司
上海安几科技有限公司	优刻得科技股份有限公司
上海物盾信息科技有限公司	杭州世平信息科技有限公司
深圳市魔方安全科技有限公司	电子科技大学
浙江孚临科技有限公司	

(以上排名不分先后)

关于研究工作组的更多介绍，请在 CSA 大中华区官网（https://c- csa.cn/research/）上查看。

在此感谢以上专家及单位。如此文有不妥当之处，敬请读者联系 CSA GCR 秘书处给与雅正！联系邮箱 research@c-csa.cn；国际云安全联盟 CSA 公众号

序言 1

在 21 世纪这个科技飞速发展的时代，人工智能技术的进步使得数字经济成为全球经济发展的主导力量。因此，各国和企业纷纷加快了对数字经济的战略规划和布局，以抢占发展制高点。

然而，在数字化转型深入进行之际，人工智能安全面临着不断加剧的挑战和风险。数据以多种形态在各个主体、不同场景下流动和应用，人工智能安全问题涉及到保密性、完整性、可用性和合规性等方面问题，同时还包括算法、模型、应用和系统的安全。传统信息安全风险管理方法已无法适应数字化业务高速发展和变化，并不能实现对人工智能全生命周期过程中风险管控。在这一背景下，《AI 安全白皮书》应运而生。本白皮书审视了 AI 在安全领域应用与挑战，并包括了 AI 赋能安全技术、伴生安全问题、监管与技术生态以及热门问题等内容，为行业提供了一个全局视角，有助于推动 AI 安全领域的进一步发展和规范。

通过对 AI 安全风险进行深入分析和研究，我们坚信企业和组织可以采取更为有效的管理措施，确保数据安全与隐私保护。面对数字经济时代，企业和组织必须高度重视人工智能安全风险管理，并积极应对各类人工智能安全挑战。借助《AI 安全白皮书》的指导和建议，我们期待广大从业者能够更好地把握人工智能安全管理要点，确保数字经济发展可持续性。同时，我们也期待更多企业和组织能够参与到人工智能安全事业中来，共同为构建安全、健康的数字经济环境贡献力量。

李雨航 Yale Li

CSA 大中华区主席兼研究院院长

序言 2

《AI 安全白皮书》是一份对当前人工智能（AI）安全领域的有指导意义的研究和分析。本文探讨了 AI 在安全领域的多种应用，包括漏洞挖掘、安全防御和威胁检测，同时也着重分析了 AI 本身的安全问题，如内生安全和衍生安全问题。此外，它还涉及了 AI 安全的监管生态、技术生态以及行业发展情况，包括法律法规、行业标准、国际共识等方面。

这份白皮书对于任何关注网络安全和人工智能发展的人来说都是有价值的资源。无论是业内专业人士、学者、政策制定者，还是对 AI 技术和网络安全感兴趣的学生和爱好者，都能从中获得独到的见解和知识。它不仅提供了对 AI 安全技术的概述，还探讨了相关的伦理法律和政策问题，为理解和应对 AI 带来的安全挑战提供了理论基础。

此外，白皮书还特别关注了 AI 安全的最新热点问题，如大模型安全、对抗样本攻击、数据投毒攻击、供应链攻击、数据泄露攻击和模型窃取攻击等，这些内容对于理解当前和未来 AI 安全的趋势至关重要。它还对 AI 伦理和 AI 辅助安全进行了讨论，为读者提供了关于如何负责任地发展和应用 AI 技术的重要思考。

总的来说，《AI 安全白皮书》是一份可读而且及时的研究成果，对于希望深入了解 AI 在安全领域应用的人来说，是不可多得的资料。它不仅涵盖了 AI 安全领域的广泛主题，还提供了关于如何面对未来挑战和机遇的有益见解。这份白皮书是理解 AI 安全领域现状和未来发展的重要参考资料。

黄连金 Ken Huang

CSA 大中华区研究院副院长

目录

背景

人工智能（AI）是一项新兴技术，代表了现代科技的发展和进步。AI 的运用范围广泛，具有广阔的扩展潜力，对各行各业领域产生了深远的影响。在数字化转型的浪潮下，AI 将成为企业成功的关键因素之一，并随着企业数字化转型的深入实践，人工智能（AI）帮助企业提升生产效率和生产力，改善客户体验和服务质量，帮助企业发现新的商机和提供更高级的安全保障。

人工智能（AI）是一项关键的革命性技术，它正在改变我们的世界。随着 AI 技术的快速发展，与数字化企业的全面实现，AI 安全问题也日益凸显。AI 安全应确保 AI 系统在执行任务时能够遵守法律、伦理规范和道德和技术的合规。在解决 AI 安全问题的过程中，应实现 AI 安全的透明度、责任化和可持续化的安全治理是企业发展的关键所在。

CSA 于 2022 年相继推出《医疗保健中的人工智能》、《ChatGPT 的安全影响》等多个安全白皮书，并进行更加全面的总结形成《AI 安全白皮书》，在 AI 安全领域通过一系列的建议和最佳实践，帮助企业遵循 AI 安全原则和标准，帮助企业解决 AI 安全的热门问题，减少潜在的风险和漏洞。以应对数字化经济时代，在确保数据可用性的前提下，创造价值，改变世界。CSA《AI 安全白皮书》也是 CSA 新兴技术认证 CSA AI 人工智能安全认证的官方学习用书，希望通过《AI 安全白皮书》的学习，对于企业全面地建立数字安全体系提供重要保障与支持。

1 AI 赋能安全

1.1 AI 赋能安全

1.1.1 AI 赋能漏洞挖掘

1.1.1.1 漏洞挖掘的发展历史

人工智能技术的发展和应用不断推动着网络安全攻防自动化和智能化的水平，目前已经在恶意代码检测、恶意流量分析、漏洞挖掘、僵尸网络检测、网络灰产检测等各领域有了广泛的应用。基于人工智能的漏洞挖掘方法通过训练大批量的漏洞代码样本，学习漏洞代码的语义、结构、指令序列等特征,从而能够自动化的定位到可疑代码。越来越多的安全公司都在增大对 AI 研究的投入，并将 AI 安全作为业务发展的重心，信息安全漏洞挖掘技术的发展历史可大致分为以下四个阶段。

第一阶段，人工代码安全审计，60 年代就开始的一种漏洞挖掘方式，由专门的安全审计人员在开发过程中或发布前后检测和评估代码。但代码安全审计是一项复杂而繁琐的任务，需要具备专业的技术和知识，投入大量的时间和人力，在安全服务领域算是工作量最大的一种服务方式，并且存在检测不全面的情况。

第二阶段，静态分析工具辅助，70 年代开始出现了一种用于分析应用程序源代码和二进制文件的技术 SAST 静态应用程序安全测试，但 SAST 无法考虑代码的上下文信息，也不能处理代码中变量和数据流的复杂性，所以有较高的漏报和误报率，所以还是需要人工进行验证和复审。

第三阶段，结合动态测试技术，90 年代开始的黑盒模糊测试就是其中一种代表性技术，通过模拟攻击者可能使用的各种手段来评估软件系统的安全性，可以更好地发现系统可能存在的安全缺陷，但漏报概率较高和覆盖率相对较低，还

是有一定的限制。

第四阶段，人工智能漏洞挖掘，近 10 年间模糊测试技术已经从传统的黑盒模糊测试技术，逐步演进到了基于覆盖引导、定向模糊测试的灰盒模糊测试技术。随着人工智能与网络安全不断交叉融合，对于 AI 赋能于漏洞挖掘也有了新的探索，AIGC 可以赋能模糊测试生成更精准的测试用例，融合覆盖引导、遗传算法、神经网络和 AIGC 等技术。

1.1.1.2 全球业界的大力推动

2018 年 3 月，美国国际战略研究中心（CSIS）在题为《美国机器智能国家战略》的报告中提出美国政府应在战略层面注重机器智能与人工智能发展齐头并进，纵观整个网络安全行业，尽管已经有很多安全公司提出了以 AI 和大数据为驱动的口号，但似乎还鲜少看到实际的应用落地。

图 1 美国 CGC 比赛现场

2023 年 9 月份美国宣布发起为期两年的人工智能网络挑战赛（AI

CyberChallenge，AIxCC），以推动自动化网络攻防技术的发展，保护美国的关键软件。该竞赛将由美国国防高级研究计划局（DARPA）牵头，旨在利用 AI 安全技术快速识别和修复关键软件漏洞。这是 DARPA 发起机器自动化网络对抗的超级挑战赛 CGC(Cyber GrandChallenge)之后，再次牵头主办的“人工智能网络挑战赛”，整个赛事的奖金总额为 1850 万美元（约合人民币 1.33 亿元），不论是从时间成本，还是资金投入，都可见美国政府对此赛事的重视程度非同一般。

放眼国内，近年来我国信息安全体系日趋完善，2017 年印发的《新一代人工智能发展规划》，标志着我国人工智能发展进入到国家战略层面。在“十四五”规划纲要中，共有 6 处提及人工智能，并将“前沿基础理论突破，专用芯片研发，深度学习框架等开源算法平台构建，学习推理与决策、图像图形、语音视频、自然语言识别处理等领域创新”视为新一代人工智能领域的重点攻关方向。

图 2 中国首届国际机器人网络安全大赛 RHG

智能化攻防技术也得到了迅速发展，RHG（Robot Hacking Game）竞赛在国内逐渐兴起，陇剑杯、黄鹤杯、强网杯、网鼎杯、纵横杯等都有 AI 攻防赛道，进一步推动人工智能技术在网络安全自动化的应用和实践。为模拟实战应用场景和检验企业的自动化安全攻防能力，这类竞赛为自动化攻防技术的探索和应用提供了实践的土壤，推动智能化攻防技术进步，标志着漏洞攻防正逐步向智能化方向演进。

1.1.1.3 漏洞挖掘的相关技术

目前，对于软件漏洞挖掘主要从源代码和二进制两个方面开展。其中针对发现漏洞的最佳实践方式有源代码审查、模糊测试、基于规则的静态分析、基于规则的动态分析等，这些方法在特定情况下针对特定二进制程序可以产生很好的检测效果，但仍然存在一些很难突破的先天缺陷，下面分别对 6 种代表性的漏洞挖掘技术作一下简要的介绍。

1.符号执行（Symbolic Execution）是一种静态分析技术，用于探索程序在不同执行路径上的行为，而不限于浅层路径。在程序执行过程中，将输入变量替换为符号变量，将程序的执行路径转化为约束条件。以达到绕过输入验证、触发故障修复或其他类型的漏洞利用的目的。

2.模糊测试（Fuzzing）用于向目标软件输入各种测试数据来发现漏洞或测试目标程序的鲁棒性。与符号执行和生成式测试的结合能够产生更多利用潜力的输入。将模糊测试中生成的输入作为符号化输入，然后进行符号执行路径探索，针对不同的输入生成一组约束条件。约束条件可以描述输入所需的特定属性或限制，例如特定的字节序列、特定的函数调用序列等。通过约束条件激活器解决这些约束条件，找到满足约束条件的具体输入，结合符号执行的约束条件和约束初始化。

3.代码插桩（Code Instrumentation）是一种修改程序源代码或二进制代码

的动态分析技术，具体取决于可用的工具和技术，以在运行时插入附加监测代码，获取代码执行过程中的指令序列，学习漏洞代码的动态执行特征。

4.控制流程图(Control Flow Graph，CFG)，针对目标二进制文件中的函数、库函数以及各种间接跳转，获得程序的控制流图的节点，结合反汇编出来的代码或者脚本语言，从而识别出可疑的汇编代码序列，进而快速有效地发现未知漏洞。

5.遗传算法（Genetic Algorithm，GA）根据大自然中生物体进化规律而设计提出的。是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型，是一种在复杂的高维空间中通过模拟自然进化过程搜索最优解的方法。该在漏洞挖掘中遗传算法可用于生成测试用例，通过迭代进化候选测试用例的群体并根据某些标准（例如覆盖率或风险）选择最佳测试用例。

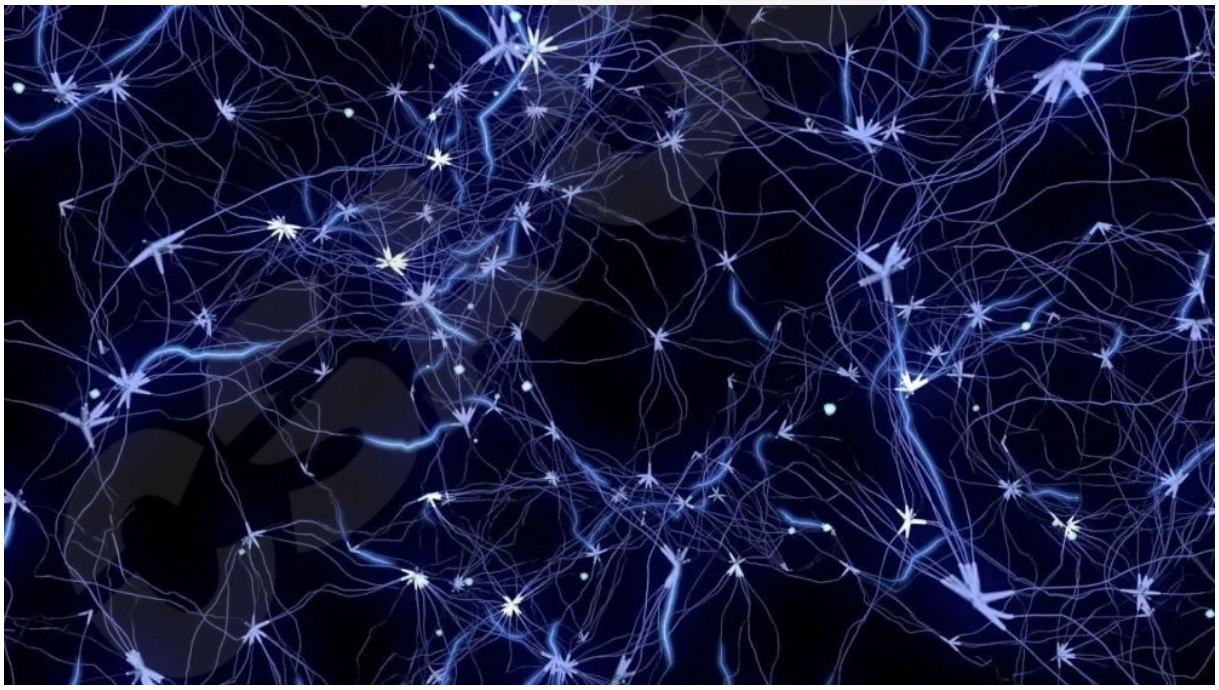


图 3 神经网络示意图（来源互联网）

6.人工神经网络（Artificial Neural Network，ANN ），从信息处理角度对人脑神经元网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络。在漏洞挖掘中，测试工具可在模糊测试过程中生成海量测试用例，以用于

训练神经网络并教育其识别海量输入中可能导致错误或其他意外行为的模式，进而形成更高效的变异策略或变异模板。随着测试数据的积累，漏洞挖掘引擎也必将愈发智能化。基于敏感函数进行代码切片，运用双向递归神经网络对代码进行训练，提取漏洞代码的静态语义特征。

1.1.1.4 自动挖掘流程与场景

通过生成式人工智能 AIGC（Artificial Intelligence Generated Content）与模糊测试相结合，模型可以根据历史漏洞数据和程序的行为特征生成更高效的测试用例，从而大幅提升测试用例质量与漏洞检测效率，自动化漏洞挖掘已经成为一种有效的技术。

首先使用模糊测试生成大量的随机输入，将这些输入提供给目标程序执行。这些输入会触发目标程序中的不同代码路径。接着利用符号执行技术对模糊测试生成的测试用例进行分析，计算程序执行的不同路径，并收集关于程序状态的符号约束信息。符号约束描述了在特定输入下程序能够到达指定路径的约束条件。从程序路径中选择部分路径，根据各个路径的约束条件，生成新的、符合路径约束的测试用例。这些生成的测试用例更有可能触发深层的未知漏洞。利用符号执行生成的测试用例验证目标程序的执行状态，检查是否出现异常行为。如果发现异常行为或异常状态，进一步分析其成因和利用可能性。根据验证结果，给模糊测试程序进行反馈，迭代模糊测试的输入生成策略，以便更好地挖掘漏洞。

然后使用代码插桩对目标程序的运行状态进行动态监测，包括代码覆盖率、路径信息和程序状态等。代码插桩为模糊测试和符号执行提供了关键的运行时数据，帮助识别异常行为和异常状态。结合模糊测试和代码插桩的结果，可以识别出目标程序中的疑似漏洞。符号执行生成的符号化输入可以帮助模糊测试触达更多路径，从而挖掘更多潜在的深层漏洞。同时符号执行也能提供精确的漏洞触发路径。插桩提供的运行时信息可以确定程序的异常行为和异常状态，从而更准确地发现漏洞。

在发现潜在漏洞后，进行验证和利用。符号执行计算出的漏洞触发路径可以精准的触达漏洞的触发点。一旦漏洞被确认，进一步构建针对该漏洞的攻击手段和攻击策略。综合利用模糊测试、符号执行和代码插桩技术，可以自动化地发现和验证目标程序中的漏洞。

自动化漏洞利用（Automatic Exploit Generation, AEG）的技术体系已经得到了广泛的关注和使用，利用计算机程序或工具自动生成针对已知或未知漏洞的利用代码或攻击创建适配器，自动生成有效的利用代码，以利用目标系统或应用程序中的漏洞，主要应用于如下几个场景：

1.可用于验证漏洞的可利用性。当一个漏洞发现并报告给厂商或开发者时，AEG 可以用于自动生成针对该漏洞的利用代码，以验证漏洞的实际利用效果，并确认其是否会导致系统被攻击。

2.用于评估漏洞修复的效果，AEG 工具重新生成利用代码，以测试修复后的系统是否仍然容易受到攻击。这种方式，可以评估修复措施的质量，并帮助开发者或厂商确保漏洞已经得到适当的修复。

3.可用于自动化的安全漏洞挖掘。它可以通过生成大量的输入和攻击适配器，探索目标系统或应用程序中的未知漏洞。通过尝试不同的输入和攻击方式，AEG 工具可以发现和生成新的漏洞利用代码，进一步帮助安全研究人员发现和修复系统中的潜在缺陷。

漏洞自动化生成利用技术是一个复杂的研究领域，涉及多种技术和方法的融合，通常情些技术是结合使用，以自动化生成针对漏洞的利用代码。

1.1.1.5 自动化挖洞未来展望

自动化漏洞大大减少了使用的难度，并且可以通过预先定义的漏洞利用模板、脚本、payloads 等工具和资源，提高了利用的速度和效率。能够覆盖多种不同

类型的漏洞，包括操作系统、应用程序、网络设备等各种目标中的漏洞。相比手动漏洞利用，自动化漏洞利用技术能够广泛地扫描和利用各种漏洞，避免由于人为疏漏或不完整扫描而导致的漏报的风险。自动化漏洞挖掘与利用已经成为红队不可或缺的一部分，提高效率，加强测试多样性，持续性的评估。不仅适用于安全研究人员，同样适用于不擅长漏洞挖掘的程序员，降低了漏洞利用环节的利用门槛。随着 AI 自我学习的完善，未来漏洞检测率将越来越高，检测成本降越来越低，漏洞检测时间将越来越短。

1.1.2 AI 赋能安全防御

随着现代社会的高度数字化，各种新型网络攻击威胁着政府、能源、金融、医疗等行业的网络安全，人工智能技术与网络安全防御融合应用是一种必然的趋势，新型网络攻击威胁不断向规模化、自动化发展，亟需打造智能化的网络安全运营体系，筑牢网络空间安全。

1.1.2.1 人工智能应用于网络安全防御的特点

网络安全已经从人人对抗、人机对抗逐渐向基于人工智能的攻防对抗发展演化，人工智能技术应用于网络安全防御领域主要有以下方面特点：

1、有效抵御复杂网络系统的安全漏洞

可通过机器学习和深度学习等方法，从大量的网络流量和日志数据中提取有用的信息，自动化、快速、准确地检测和修复复杂网络系统中的安全漏洞，从而更快发现并解决安全漏洞，提升整体安全防御水平。

2、大幅缩短对攻击的响应时间

通过人工智能技术可以在短时间内快速分析网络流量和日志数据，从而在攻击威胁发生时能够实时检测并响应。与传统安全防御技术相比，人工智能技术可更快实现网络自主监测，发现攻击威胁，大幅缩短响应时间，降低误报率，减少

潜在损失，提高网络安全防御的效能。

3、增强网络安全防御的协作能力

通过人工智能技术应用帮助网络安全运营团队提升应急响应协作的效率，实现团队之间高效的信息交流、共享和写作，提高整体的安全防御效果赢得网络安全防御主动权，最大限度阻断威胁、降低风险。

1.1.2.2 人工智能在网络安全防御中的应用

人工智能技术应用可以提升网络安全防御的智能化水平，在网络安全防御中的应用包括以下几个方面：

1、智能安全漏洞防御

传统的安全漏洞检测主要依靠人工的方式进行，专家依赖性强且耗时长，通过人工智能技术实现自动化漏洞发现与修复，极大提升安全漏防御能力。

- **安全漏洞发现：**

可通过机器学习算法分析网络流量、系统日志、攻击特征等多维数据，识别漏洞特征，通过深度学习准确地判断是否存在漏洞。

结合神经网络和自然语言处理等人工智能技术，识别并解释代码语法含义以评估风险，分析漏洞利用趋势，缩小代码审计范围，减少开发人员检测和发现漏洞的时间。

- **安全漏洞修复：**

由于漏洞类型繁多、漏洞定位困难等因素的存在，修复漏洞需要大量人工参与，自动化漏洞修复技术能够极大提升漏洞修复效能。

机器学习算法可以通过分析漏洞特征和历史修复方案，自动学习漏洞修复方

法，并为工程师提供修复建议。通过对漏洞的深度分析和模拟测试，自动生成修复代码或提供修复工具，帮助工程师快速修复漏洞，提升漏洞修复效率。

2、智能安全态势感知

随着网络部署规模复杂化，网络安全情报数据种类和数量激增，对网络安全态势感知提出更高要求。人工智能技术在安全态势感知中的应用，可以提高数据处理效率，增强关联分析能力，自动检测和响应威胁，打造智能安全防御体系，提高网络安全性：

- **提升海量数据处理效率：**

人工智能技术可以自动化地收集和处理大量的网络安全威胁情报数据，包括来自安全设备、网络流量、安全事件等来源海量数据，高效地完成数据处理，快速、准确地识别出威胁。

- **增强威胁数据挖掘能力：**

通过学习和分析大量网络安全威胁数据，识别出威胁模式和攻击者的行为特征，将多个独立的安全事件进行关联和分析，挖掘潜在关系，多维构建攻击者画像、建立威胁情报库，更全面掌握高级威胁情报。

- **智能检测和预警：**

通过对威胁情报数据智能分析，实现自动检测和识别潜在的安全威胁，通过对历史数据的分析和预测，预警未来可能出现的威胁趋势和攻击模式。

3、智能安全运营

网络安全运营面临诸多挑战，例如安全设备种类繁多、威胁数据量庞大、专业人才短缺以及人工处理效率低下等问题。将人工智能技术与安全运营相结合，能够显著提升安全运营的智能化水平，极大提高处理效率。

AISecOps（智能驱动安全运营）是人工智能技术与安全运营的融合，以安全运营目标为导向，以人、流程、技术与数据的融合为基础，面向预防、检测、响应、预测、恢复等网络安全风险控制环节，构建具有高自动化水平的可信任安全智能模型。通过有效纳管企业安全产品，实现模块整合、统一运营、指标量化，精准判断安全事件影响范围，快速指挥调度及响应。

人工智能技术应用可以从以下几方面提升安全运营效能：

- **智能化响应与处置：**

将威胁模型实践与实时网络流量相关联，通过自然语言处理 NLP、语义分析、上下文处理，并结合正则匹配、关联分析和迁移学习等技术，能够准确实现智能分析研判，加速安全策略自动下发提升安全响应速度和效率，减少安全事件的影响范围和损失。

- **安全运营流程智能化：**

打造人机智能协同的算法、模型、系统与流程，才能不断地适应高级别的智能化安全运营场景，实现安全运营全流程智能化，包括安全审计、漏洞扫描、合规性检查、安全事件处理等。

- **智能分析预测与决策：**

可以通过数据分析和模式识别，预测未来的安全趋势和威胁。可以为决策者提供重要信息，优化安全运营战略和决策。

- **自动化生成合规性报告：**

支持自动生成安全运营报告，满足合规性要求。通过人工智能技术自动化分析整体安全态势，生成多维度、安全运营报告。

- **打造智能安全防御体系：**

通过与其他安全设备结合和联动，如防火墙、入侵检测等，实现不同安全防御设备联动调度，打造动态的、智能的安全防御体系。通过持续监控和评估安全防御系统的效果，优化模型和算法，提高智能安全检测的准确度和效率，提高网络安全整体水平。

4、智能网络攻击溯源

随着网络攻击和数据泄露等威胁不断增加，在网络安全防御中，网络攻击溯源扮演着重要角色，网络攻击溯源结合人工智能技术，能够提供更准确、更智能的网络安全保护。

网络攻击溯源通过安全设备告警、日志和流量分析、服务 资源异常、蜜罐系统等对网络攻击进行捕获，发现威胁，利用已有的 IP 定位、恶意样本分析、ID 追踪、溯源反制等技术，收集攻击者信息。通过对攻击路径的绘制和攻击者身份信息的归类形成攻击者画像，完成整个网络攻击的溯源。

1.1.2.3 人工智能在网络安全防御中的挑战

1、网络安全场景攻击趋于复杂化，基于已知样本集预测未知难度较大

人工智能算法本质是基于样本抽象出特征，进而基于特征表征问题，用于新数据推理判定。但网络安全领域中标注样本少，攻击方式多样化、复杂化，仅基于已知样本构建的人工智能算法模型难以覆盖所有攻击场景。

2、算法模型如果不能自适应优化，可能导致无效告警产生

通过人工智能算法可以发现更多的安全问题，但不同客户场景业务、网络和资产属性不尽相同，会导致通用模型检测出的告警中存在一定的无效告警或误报，如果算法模型不能自适应优化，需要安全专家重复研判分析，这将大大增加人工成本。

3、人工智能算法模型检出告警的可解释性需要关注

如果告警的可解释性足够高，安全专家才能从事件描述中了解攻击详情，对威胁进行精准溯源和处置。基于专家规则或 IOC 检测的告警，解释性较高。但人工智能算法模型自身缺乏透明度，若其检出的告警描述仅呈现如报文大小、概率值等特征原始数值，将很难支撑安全专家进行响应处置。

1.1.3 AI 赋能威胁检测

从广义上来说，符号逻辑、规则与专家系统等都在 AI 的范畴内，那么 AI 赋能网络安全领域的威胁检测应用可能已有 40 多年或更长的历史；即使是逻辑回归、支持向量机、贝叶斯等机器学习算法的应用，也在上个世纪 90 年代就已出现。受限于篇幅，同时也从时效性的角度出发，本节将对近些年来更为流行的 AI 赋能威胁检测应用的相关技术给出简要介绍。

1.1.3.1 恶意代码检测

网络空间时刻面临着以“僵木蠕”（僵尸网络、木马、蠕虫）攻击为典型代表的严峻安全威胁，而恶意代码通常是这些攻击的重要载体，包括近些年来在全球范围内广为流行的勒索病毒、挖矿病毒等，也往往由某种恶意代码承载。因此，精准有效的检测恶意代码对于网络安全保障有着不可忽视的作用。由于文件类型和系统平台的多样性，恶意代码也有很多种类，但检测技术上都存在一定的相似性。不失一般性，本小节以最常见的恶意代码类型之一——二进制 PE 文件为例，说明 AI 如何赋能恶意代码检测。从总体上看，经典的检测方法可以分为两大类。

- 基于静态分析的方法：直接对恶意代码本身进行分析，从中提取可用于识别的特征签名（如病毒的特征码等）作为识别的直接依据，也可对属于同一类别的多个恶意代码特征进一步总结出识别的规则。

- 基于动态分析的方法：在虚拟环境中运行恶意代码并监控其行为，如系

统调用、文件系统访问、网络传输等，从中找出异常及可能造成危害的行为模式，并作为识别的依据。

实际上，这两类方法均可采用不同的实现方式，除了上面所描述的基于特征签名和基于规则的方式外，也都可以采用机器学习、深度学习等 AI 技术，即基于恶意代码的多维度特征构建智能检测分类模型。相比来说，AI 技术的应用能够有效提高检测分类的自动化水平，减少了人工定义检测规则或特征的工作量，而且 AI 模型可以通过使用新样本进行迭代训练而实现自动更新优化，避免了规则或特征库更新所需的人工投入。

1.1.3.2 恶意流量检测

网络攻击总是需要通过网络流量来承载，如何能高效精准识别这些承载攻击的恶意流量一直是网络安全领域的重要研究问题。随着技术的发展，恶意流量越来越呈现出隐蔽性和多样化的特点，传统的基于特征指纹的检测手段难以达到安全防护的目标，尤其是如今加密流量占比日益增加，更是给恶意流量的检测识别带来了新的挑战。AI 技术在恶意流量检测中的应用也很广泛，具体有如下两种典型方式。

（1）基于历史流量数据构建异常检测模型。这类方法相对较为简单，但也应用广泛。一方面，可以采用不同的 AI 技术构建统计基线以用于异常检测，包括面向单数值特征的数值分布区间、正态分布拟合及 3σ 准则等，面向时间序列的 ARIMA、指数平滑等，面向多维度特征的 One-Class SVM、自编码器等。另一方面，还可以选择一些不需要预训练而直接使用的异常检测算法，包括孤立森林、ABOD 等。

（2）针对流量数据提取多维度特征构建基于机器学习的分类检测模型。其中，根据特征来源不同，可分为单包特征、统计特征、包序列特征等。具体来说，一些恶意攻击的特征直接体现在对应的单个报文里，因此可以提取一系列单包特

征用于构建分类检测模型，例如 HTTP 协议头中的 URL、referer、user-agent、cookie 等字段特征，报文净荷中包含的明文敏感函数调用，加密流量中使用的证书特征等。另一方面，针对单包和流会话的相关统计特征也可用于区分正常流量与特定类型的恶意流量，如信息熵、可见字符占比、参数个数及意义分值、base64 编码长度等。此外，流会话中的上下行包大小序列及包时间间隔序列相关特征也广泛用于构建识别加密恶意流量的检测模型中。在完成相应特征提取及特征工程后，通常采用逻辑回归、LGBM、xgboost 等机器学习算法进行建模实现对恶意流量的有效检测分类。

（3）针对流量数据直接构建基于深度学习的分类检测模型。类似于恶意代码检测中将二进制 PE 文件当作灰度图像分类的思想，网络流量数据也可以转换成灰度图像的样式，然后采取 CNN 等深度学习算法进行建模而无需进行单独的特征计算提取步骤。另一方面，上面所提到的包大小序列和包时间间隔序列也可以采用 RNN、LSTM、Transformer 等深度学习算法直接进行建模。

1.1.3.3 恶意域名检测

恶意域名有不同的种类，有的是对应网站包含不良信息或恶意代码，有的是仿冒的钓鱼网站，也有的是 DGA 算法生成的随机性域名用于建立 C&C 通信，等等。对于这些不同类型的恶意域名，检测方法与所需要的数据也是不同的。单纯使用域名黑名单的方法无法识别新出现的或未知的恶意域名，而且恶意域名通常也无法使用正则表达式等规则匹配的方式进行精准识别。本小节对一些 AI 赋能恶意域名检测的常见做法给出简要描述。

（1）基于域名字符串的检测方法

对于 DGA 算法生成的恶意域名，或者仿冒的钓鱼网站域名等场景，由于其域名本身包含较为丰富的特征，所以可基于域名字符串提取一系列特征以构建 AI 模型。因为域名是文本字符串格式的，所以可借鉴采用文本分类中的相关技术，

从简单的 *n*-gram 特征提取，到 LSTM、Transformer 等深度学习模型的应用，技术细节这里不再赘述。

（2）基于域名请求解析记录的检测方法

除了域名字符串本身外，针对域名的请求解析记录数据能够为恶意域名的检测判定提供更多的信息。通过对解析记录数据提取多维度特征，包括域名解析的结果、解析请求的时间及 IP 分布等，基于有监督机器学习算法可以构建更精准的恶意域名检测分类模型。

（3）基于网站内容的检测方法

有些恶意域名需要结合对应网站的页面内容进行更精准的判定，例如包含不良信息或植入恶意代码的网站、仿冒的钓鱼网站等。通过爬取网页并从中提取更多特征，包括页面的文本内容及图片元素、内嵌的 JS 代码等，然后构建相应的检测分类模型。

1.1.3.4 威胁情报提取

威胁情报对于网络安全威胁检测与防护有着非常重要的价值，但是很多原始情报信息存在于非结构化的文本中，比如攻击事件报告、漏洞信息描述等，需要经过提取才能成为更易用的结构化情报，而且由于自然语言的灵活性，通常无法使用基于正则表达式等模板匹配方式完成自动化提取。下面以从漏洞信息中提取有价值的情报场景为例，说明相关 AI 技术是如何应用的。

很多来源的漏洞描述信息是以非结构化文本形式呈现的，包括 CVE、ExploitDB、SecurityFocus、SecurityTracker、Openwall、SecurityFocus 等。从这种非结构化文本中提取有价值的威胁情报，如受漏洞影响的软件名称和版本等，面临多项技术挑战：（1）由于软件名称和软件版本等相关情报信息的多样性，基于字典和正则表达式的方案难以实现高精确率和高召回率；（2）漏洞描述信息的非结构化文本经常包含代码，而且其独特的写作风格也使得传统的自然

语言处理算法难以驾驭；（3）需要在提取有价值情报的同时排除文本中无关的实体，如不受漏洞影响的软件名称及其版本。鉴于上述原因，通常采用基于深度学习模型的自然语言处理技术来完成威胁情报的自动化提取的。具体来说，首先需要使用命名实体识别模型来识别感兴趣的实体（即期望提取的情报元素），例如受漏洞影响的软件名称及版本，然后使用关系抽取模型将识别出的实体进行配对，例如漏洞可能影响到多个软件的若干版本，则需要将软件名与对应的版本信息配对。其中，上述命名实体识别模型可采用 BiLSTM、BiGRU、BERT、UIE 等，上述关系抽取模型可采用 HAN、CasRel、TPLinker、UIE 等。

1.1.3.5 敏感信息识别

随着互联网的飞速发展，网络成为了信息传播的重要渠道，同时内容安全、信息泄露等问题日益突出。为了避免敏感信息的外泄对个人、企事业单位、甚至国家安全和利益造成的威胁，必须对敏感信息与数据进行严格管控。因此，如何高效精准地识别敏感信息成为一个需要解决的重要问题。敏感信息的表现形式多样，其中最常见的一种就是以文本内容为主的敏感文档。

传统的敏感文档识别技术主要基于关键词表与词频统计，将文档中是否出现关键词及出现数量作为主要识别依据。然而，现实存在另一类比较广泛的应用场景却不适用这一方法。在这类场景中，会预先指定一批文档为敏感文档，需要识别与指定文档语义相近的所有文档。注意这些指定文档未必是一般意义上的敏感文档，可能不包含特定的敏感词语，比如内部会议纪要等。

针对上述问题，核心关键是计算不同文档之间的相似度，基于与指定文档相似度的高低来判定是否敏感。为了快速计算大量文档间的相似度，可以先采用词嵌入向量、文档嵌入向量等 AI 技术将一篇文档表示为多维向量，然后使用余弦距离、词移距离等距离度量来作为相似度的表征，再运用聚类方法将所有文档自动聚成不同的分簇，并将与指定敏感文档归为同一分簇的文档标记为敏感，从而实现一次性对大量文档完成标记工作。

另一方面，随着社交网络平台的发展，用户发布的信息通常由图片、文本等多种格式组成，仅从文本或图片单一维度进行敏感信息识别是不全面的，而必须采用多模态方法。一种简单而实用的思想是分别针对文本与图像构建相应的敏感信息识别分类模型，然后再在决策层使用融合算法进行最终判定。

1.2 AI 伴生安全

1.2.1 概述

新技术解决问题的同时会带来安全问题，这是新技术在安全方面的伴生效应。这种伴生效应会产生两方面的安全问题：一是由于新技术的出现，其自身的脆弱性会导致新技术系统出现不稳定或者不安全的情况，这方面的问题是新技术的内生安全问题；另一方面是新技术的自身缺陷可能并不影响自身的运行，但这种缺陷却给其他领域带来问题和风险，这就是新技术的衍生安全问题。人们首先会去享用新技术带来的红利，之后才会注意到新技术伴随的种种网络安全问题，比如先有云计算，后有云计算安全。同样，人工智能作为一项新技术革命，既能赋能安全，又具有伴生安全问题。

目前， AI 的安全问题伴生效应也会表现在两个方面：一个是 AI 系统的脆弱性导致自身出现问题，无法达到预设的功能目标，即自身的内生安全。从人工智能内部视角看，人工智能系统和一般信息系统和应用一样，由软硬件系统组成，会存在脆弱性，一旦人工智能系统的脆弱性被恶意分子利用，就可能引发安全事故，还有一种情况是新技术存在着天然的缺陷，比如人工智能依赖于数据与算法，数据的丢失和变形、噪声数据的输入、数据投毒攻击都会对人工智能系统形成严重的干扰。

伴生效应的另一个方面是系统的脆弱性被攻击者所利用或不恰当地使用，从而引发其他领域的安全问题。比如，人工智能系统的不可解释性也许不会影响人工智能系统在一般情况下的运行状态，但可能会导致人工智能系统可能出现失控

的情况，造成系统可能会不按照人类所预计的或所期望的方式运行，而这种不可预计的情况可能给人类带来威胁。比如，如果一个养老机器人不按照人类期望的方式运动，可能会伤害人类。还有一类具有移动性、破坏力、可自主学习的人工智能系统（比如机器人）有可能从人类为其设定的约束条件中逃逸，进而危及人类安全，这些都是人工智能的衍生安全问题。

另外，随着 ChatGPT 等大模型的广泛使用，某些别有用心的人将其作为违法活动的工具。例如生成虚假新闻、深度合成伪造内容进行诈骗或钓鱼攻击，侵犯他人肖像权、隐私圈，输入的语料和生产的内容也可能涉及知识产权方面的纠纷。

AI 衍生安全影响 AI 的合规使用，还涉及人身安全、隐私保护、军事与国家安全、伦理道德和法律规范等一系列与社会治理有关的挑战性问题。

1.2.2 AI 内生安全

人工智能内生安全指 AI 技术的脆弱性、对数据的依赖性等自身缺陷原因而带来的问题。

AI 技术近几年的快速发展依赖于深度学习算法、算力和数据。针对 AI 的威胁在总体上分为三个大方面，即 AI 模型和算法、数据与 AI 依托的信息系统，后者指软硬件环境和网络等。

AI 内生安全涉及多个方面：

- 人工智能依赖的框架、组件、环境等存在问题
- 人工智能数据的噪声、不平衡、错误等问题
- 人工智能算法缺陷问题
- 人工智能模型的知识产权保护、污染等问题

2023 年， OWASP 发布的机器学习安全风险 TOP10 就涵盖诸多方面。

ML01:2023	对抗性攻击 Adversarial Attack	当攻击者故意更改输入数据以误导模型时，就会发生对抗性攻击。
ML02:2023	数据投毒攻击 Data Poisoning Attack	当攻击者操纵训练数据导致模型以不良方式运行时，就会发生数据投毒攻击。
ML03:2023	模型反转攻击 Model Inversion Attack	当攻击者对模型进行逆向工程以从中提取信息时，就会发生模型反转攻击。
ML04:2023	成员推理攻击 Membership Inference Attack	当攻击者操纵模型的训练数据以使其行为暴露敏感信息时，就会发生成员推理攻击。
ML05:2023	模型窃取 Model Stealing	当攻击者获得对模型参数的访问权时，就会发生模型窃取攻击。
ML06:2023	损坏的组件包 Corrupted Packages	当攻击者修改或替换系统使用的机器学习库或模型时，就会发生损坏的组件包攻击。
ML07:2023	迁移学习攻击 Transfer Learning Attack	当攻击者在一个任务上训练模型，然后在另一个任务中对其进行微调，导致结果未按照预期产生，就会发生迁移学习攻击。
ML08:2023	模型偏斜 Model Skewing	当攻击者操纵训练数据的分布，导致模型以不希望的方式运行时，就会发生模型偏斜攻击。
ML09:2023	输出结果完整性攻击 Output Integrity Attack	当攻击者的目的是为了改变其 ML 模型的行为或对使用该模型的系统造成损害，从而修改或操纵 ML 模型的输出结果，就会发生输出结果完整性攻击。
ML10:2023	神经网络重编程 Neural Net Reprogramming	当攻击者操纵模型的参数使其以不良的方式运行时，就会发生神经网络重编程攻击。

图 4 OWASP 机器学习安全风险 TOP10

框架（如 TensorFlow、Caffe）是开发人工智能系统的基础环境，重要性不言而喻。当前，国际上已经推出了大量的开源人工智能框架和组件，并得到了广泛使用。然而，由于这些框架和组件未经充分安全评测，可能存在漏洞甚至后门

等风险。

训练后投入使用的模型是机构重要的资产，面临被窃取或污染的风险，需要有知识产权保护和风险防范与检测机制。

在算法方面，难以保证算法的正确性是人工智能面临的一大问题。现在的智能算法普遍采用机器学习的方法，直接让系统面对真实可信的数据来进行学习，以生成机器可重复处理的形态，在可靠性、公平性、可解释性、透明性和鲁棒性方面存在安全缺陷。例如，对抗样本就是一种利用算法缺陷实施攻击的技术，自动驾驶汽车的许多安全事故也可归结为由于算法不成熟而导致的。

以深度学习为代表的人工智能技术与数据是相辅相成的，数据安全是人工智能安全的关键要素，人工智能系统高度依赖数据获取的正确性。然而，数据正确的假定是不成立的，有多种原因使得获取的数据质量低下。例如，数据的丢失和变形、噪声数据的输入，都会对人工智能系统形成严重的干扰。

人工智能在部署后才能应用，可能由于主客观原因导致运行时出现安全问题。

下面分布对 AI 框架、算法、模型、数据和运行五个方面进行说明。

1.2.2.1 框架安全

人工智能算法从设计到实现需要经过很多复杂的过程，许多程序在不同的算法中是可以高度复用的，因此出现了很多深度学习的框架，提供了常用的函数和功能等，供开发者以更简单的方式实现人工智能算法。通过使用深度学习框架，算法人员可以无需关心神经网络和训练过程的实现细节，更多关注应用本身的业务逻辑，目前使用较多的深度学习框架包括 TensorFlow、PyTorch、PaddlePaddle 等。

此外，算法模型和框架还强依赖于大量三方包，如 numpy、pandas、计算机

视觉常用的 openCV、自然语言处理常用的 NLTK 等等。因此，一旦这些深度学习框架和三方包中存在漏洞，就会被引入模型，并破坏模型的可用性。

例如，CVE 公布的 CVE-2019-9635 漏洞指出 TensorFlow 1.12.2 之前的版本存在 “空指针解引用” 漏洞，可通过构造特殊 GIF 文件对系统进行 “拒绝服务攻击”。这几年，关于 AI 框架与组件的漏洞急速提升，各主流平台无一幸免，漏洞类型涵盖缓冲区溢出、CSRF、XSS 等。

1.2.2.2 算法安全

算法是人工智能系统的大脑，定义了其智能行为的模式与效力。

现在的智能算法普遍采用机器学习的方法，直接让系统面对真实可信的数据来进行学习，以生成机器可重复处理的形态，最经典的当属神经网络与知识图谱。神经网络是通过“输入-输出”对来学习已知的因果关系，通过神经网络的隐含层来记录所有已学习过的因果关系经过综合评定后所得的普适条件，神经网络在很多问题上均能取得十分优异的表现，但是神经网络为什么能取得如此好的效果，神经网络中的多个隐藏层分别代表什么含义，神经元的参数等是否具有具体的意义?这些问题目前都很难回答，神经网络如同是一个黑盒子，具有不可解释性。

可解释性要求强的场景中，神经网络造成的失误则会造成巨大的损失。例如，在医学场景中，如果让智能算法自动识别某患者是否患病，神经网络不具有可解释性，仅会给出患者是否患病的分类结果；如果算法错误，则影响很大，在这样的场景中，不论是病人还是医生都希望这样的智能算法能给出解释，即是什么症状及检查结果让算法识别为患病。

综合来说，以机器学习为代表的算法自身在可靠性、公平性、可解释性、透明性和鲁棒性方面存在安全缺陷，决策过程如同黑盒不可预见。从外部威胁角度来看，对抗样本攻击、数据投毒等行为使得智能系统产生错误的分类模式，逆向

攻击技术可以通过大量的模型预测查询实现模型窃取。对抗样本就是人工智能算法缺乏可解释性的一种体现，自动驾驶汽车的许多安全事故也可归结为由于算法不成熟而导致的。

1.2.2.3 模型安全

通过大量样本数据对特定的算法进行训练，可获得满足需求的一组参数，将特定算法和训练得出的参数整合起来就是一个特定的人工智能模型。可以说模型是算法和参数的载体并常以实体文件的形态存在，模型也是组织非常重要的资产，对其知识产权的保护也成为挑战。

例如，TensorFlow 框架通过 Saver 对象将这些参数保存为.ckpt 文件，当我们使用 TensorFlow 机器学习框架时，攻击者就可以根据这些框架默认的保存路径、保存方式等获取我们学习到的模型，从而实现对模型的窃取。另外，在一些云边协同场景中模型参数的传递、边缘端设备的存储和管理能力差等也会导致模型的安全问题。例如，当攻击者攻击了云端服务器或边缘端设备，利用漏洞便可以获取模型参数，而这些模型参数的意义巨大，一旦丢失会造成严重的后果。

另外，随着人工智能算法的普及，很多开发者愿意将训练后的模型共享给更多用户，攻击者便有机会对开源的模型进行攻击，通过下载或购买此类模，人工智能模型的共享也将会越发普及，通过模型水印等技术保护知识产权，并采取手段防止模型被非法窃取。

1.2.2.4 数据安全

数据是人工智能的重要基础，深度学习掀起人工智能发展的又一热潮，其中一个重要的原因便是近十几年来大数据的蓬勃发展为机器学习等人工智能算法提供了大量的学习样本，使人工智能相关技术迅速发展。在此前提下，数据安全就成为了人工智能内生安全的重要部分，具体来说，数据集质量、数据投毒、对

抗样本都可以影响人工智能的安全。

数据集的质量取决于数据集的规模、均衡性、准确性等，有多种原因使得获取的数据质量低下，如数据的丢失和变形、噪声数据的输入，都会对人工智能系统形成严重的干扰，直接影响到人工智能算法的执行效果。

数据集的质量能影响人工智能算法的安全性，人为地对数据进行修改则能在很大程度上改变人工智能算法的执行效果。比如，有些人工智能算法为了实时地适应数据分布的变化，会周期性地采集近期历史样本数据以重新训练人工智能模型参数。如果攻击者掌握了人工智能算法周期性采集数据的规律，就可以对即将被采集到的数据进行污染，让人工智能算法学习到错误的数据特征，从而歧化人工智能算法的模型参数，造成算法失效，这种攻击方式被称作数据投毒。数据投毒可以说是对人工智能算法训练过程进行的攻击，即通过输入不正确的样本数据，使得人工智能算法训练得到不正确的模型参数，从而引发算法错误。

相比于数据投毒，通过对抗样本实施攻击是近年来新出现的一种攻击方法。这种方法在不改变模型参数的情况下，对人工智能算法需要识别的数据加以修改，让算法失效，是对人工智能算法识别过程的攻击。从攻击方式上来划分，采用对抗样本的攻击方式可分为白盒攻击和黑盒攻击；从攻击效果上来划分，可分为无目标攻击和有目标攻击；从对抗样本的形式上来划分，可分为针对图像、文本、音视频的攻击。

1.2.2.5 运行安全

人工智能模型需要进行实际部署才能应用，而在部署以后可能由于客观或主观的原因导致人工智能模型运行时出现安全问题。

人工智能应用依托算法模型、数据和算力基础设施构建而成，面临广泛的攻击面，系统在应用过程中涵盖开发、测试、配置、部署、使用、数据处理、存储

等完整信息系统生命周期，与框架安全、算法安全、模型安全、数据安全和基础设施安全高度相关，如果算法所在的环境发生了变化，而经过抽象的数据如果不能描述这种变化，人工智能模型则无法分辨环境改变对数据和算法造成的改变。

再比如如果攻击者利用了框架漏洞，不仅能盗取模型参数，更严重的后果将是利用模型实施恶意的攻击行为，如让人脸识别模型不能正确识别人，让智能家居做出异常举动造成恐慌等。因此，应从客观上尽量保证人工智能模型的安全性，从数据、框架、算法、模型、基础环境和应用等多个层面对人工智能模型进行安全检测，在整个运行生命周期重从技术和管理角度及时发现模型漏洞与不足。

1.2.3 AI 衍生安全

人工智能衍生安全指人工智能系统因其自身脆弱性被利用而引起其他领域的问题，AI 衍生安全影响 AI 的合规使用，还涉及人身安全、隐私保护等。

总体来说，人工智能的衍生安全包括以下两大类：

- 人工智能使用的不正当：应用目的产生危害，如应用于武器、欺骗人脸支付等方面。
- 人工智能被不正当地使用：应用过程存在伦理性问题，如自动驾驶刹车时是保护司机还是乘坐者。

衍生安全的范围很广，下面仅从人工智能系统失误而引发的安全事故、大模型带来的安全问题和人工智能武器引发军备竞赛等三个方面简要介绍衍生安全的挑战。

1、人工智能系统失误而引发的安全事故

人工智能系统（如机器人）一旦同时具有行为能力以及破坏力、不可解释的决策能力、可进化成自主系统的进化能力这 3 个失控要素，不排除其脱离人类控

制和危及人类安全的可能。

据美国网络安全公司 IOActive 对 50 个机器人进行了安全调查，发现 10 个机器人中有近 50 个安全漏洞可能威胁到人身安全，如果人工智能系统经常犯错，基于人工智能的物联网体系就会变成令人感到恐怖的系统。再比如由于特斯拉自动驾驶汽车使用许多传感器，这些传感器不断向自动驾驶系统发送数据，因此恶意攻击者可以通过攻击数据源、采用数据欺骗或其他手段来远程控制汽车系统，由此可导致汽车偏航或重大交通事故。

2、大模型广泛应用带来的衍生安全问题

随着 ChatGPT 等大模型技术的成熟，生成式人工智能服务被广泛地应用，会带来一系列的安全问题。大模型在数据集的搜寻上有可能存在价值观的偏见，从而形成不同意识形态的传播问题，生成的内容也可能带有歧视、偏见、甚至对他人或社会造成危害。利用生成式大模型生成的作品可能包含一些伪造的内容，被用于诈骗与钓鱼攻击，也可能侵犯他人肖像权、隐私权、名誉权，假如生成虚假新闻则可能引发更加复杂的社会问题。AIGC 用到的语料和生成的内容可能涉及侵犯知识产权方面的问题，也可能泄露隐私和敏感数据。

3、人工智能武器研发可能引发国际军备竞赛

将发展人工智能列入国家战略已成为很多国家的共识。在美国确定将发展人工智能技术作为其核心战略之后，很多国家纷纷效仿，努力实现从开发到应用的跨越式进步，一场人工智能技术军事化应用的竞赛激烈展开。这些积极发展人工智能武器的国家大多以巩固国家安全为借口或目标。早在 2016 年，美国已就人工智能对国家安全问题的影响发布了多个白皮书，从国家军事部门到武器开发商，都在为打赢未来战争积极准备。进入 2018 年，美国为人工智能武器化发展做了更加频繁的实质性推进工作，包括但不限于：使无人机逐步具有自主作战能力，将人从军事决策中解放出来、使人工智能武器的定位功能摆脱对卫星定位系统的依赖；自动射击机器人等智能武器；谋划在未来战争中打造出完全自主化的智能

武器，以算法精准的智能武器打造美军的战场制胜能力；计划运用人工智能技术有效提升战场军人的体能素质和作战能力，打造形成钢筋铁骨的“人机”部队，强化美军的战争能力。

因人工智能系统失误而引发的安全事故、人工智能武器研发可能引发国际军备竞赛这两类属于实际发生的人工智能衍生安全问题，随着人工智能的广泛应用和快速发展，人们深深担忧人工智能未来可能会失控，会带来更多的衍生安全问题，AI 内生安全问题的发展将会会导致 AI 衍生安全问题的多样化。

2 AI 安全的生态

2.1 AI 安全的监管生态

2.1.1 法律法规

美国白宫 2023 年 10 月 30 日发布拜登签署的《关于安全、可靠和可信的 AI 行政命令》，以确保美国在把握 AI 的前景和管理其风险方面处于领先地位。作为美政府负责任创新综合战略的一部分，该行政令以美国总统之前采取的行动为基础，包括促使 15 家领军企业自愿承诺推动安全、可靠和可信的 AI 发展的工作。该行政令包含 8 个目标：

（1）建立 AI 安全的新标准；

（2）保护美国民众的隐私；

（3）促进公平和公民权利；

（4）维护消费者、病患和学生的权益；

（5）支持劳动者；

（6）促进创新和竞争；

（7）提升美国在海外的领导力；

（8）确保美国政府负责任且有效地使用 AI。

美国《2020 年国家人工智能倡议法案》（National AI Initiative Act of 2020）颁布于 2021 年 1 月，旨在强化和协调各联邦机构之间的人工智能研发活动，确保美国在全球人工智能技术领域的领先地位。《2020 年国家 AI 倡议法案》通过将美国 AI 计划编入法典以帮助增加研究投资、改善计算和数据资源的获取、设置技术标准、建立劳动力系统并跟盟友展开合作。其中关键措施有：

- 设立国家人工智能倡议办公室，属于白宫科技政策办公室，承担监督和实施美国国家人工智能战略等职责。

- 设立国家人工智能咨询委员会，委员会应由商务部部长任命，代表广泛跨学科的学术机构、私营企业、非盈利机构等，并就人工智能相关事项向总统和国家人工智能办公警示提供建议，帮助美国保持在人工智能领域的领导地位。

- 加大人工智能研发投入，命令联邦机构在其研发任务中优先考虑人工智能投资的方式，保持美国对高回报、基础性人工智能研发的长期且强有力的重视。

- 开放人工智能资源，要求相关机构将联邦数据、模型向美国人工智能和计算资源研发专家、研究人员和产业开放，增强公众对人工智能技术的信任，提高这些资源对人工智能研发专家的价值，同时确保数据安全、保护公民自由与隐私权又不失机密性。

- 设定人工智能治理标准。联邦机构将通过建立适用于不同领域的技术工业部门的人工智能发展指南来增强公众对人工智能系统的信任，帮助联邦监管机构制定一套人工智能技术的治理方法。该倡议还要求美国

国家标准与技术研究所领导制定人工智能系统的适当技术标准，使其可靠、安全、便捷、可互操作。

依据此法案，美国白宫科学技术政策办公室（OSTP）宣布成立国家人工智能倡议办公室和国家人工智能咨询委员会，并建立或指定一个机构间委员会，以更健全完备的组织机构推动“国家人工智能计划”实施。该法案显示了两党对美国政府在 AI 领域长期努力的大力支持，并将美国政府现有的许多 AI 政策和举措编入法律并加以扩展。例如，将美国 AI 计划确立的 5 项关键任务（加大人工智能研发投入、开放人工智能资源、设定人工智能治理标准、培养人工智能劳动力，以及国际协作和保护美国人工智能优势）纳入法律，扩大 2018 年成立的人工智能专责委员会并使其成为常设机构，承认 2020 年成立的国家 AI 研究所的合法地位，规定要对 2019 年发布的国家 AI 研发战略规划进行定期更新，将白宫 2019 年指导的关键 AI 技术标准活动扩展至包括 AI 风险评估框架等。

美国《人工智能权利法案蓝图》

2022 年 10 月美国白宫科技政策办公室发布了《人工智能权利法案蓝图：让自动化系统为美国人民服务》。该文件目的是支持制定政策，在自动化系统的建设、部署和治理中保护公民的权利和促进民主价值。核心内容如下：

- 安全和有效的系统。自动化系统应该通过咨询不同的社区、利益相关者和领域专家，以确定系统的关注点、风险和潜在影响。

- 算法歧视的保护。系统应该以公平的方式使用和设计，当自动化系统导致人们因其种族、肤色、民族、性别（包括怀孕、性别认同等）、宗教、年龄、国籍、残疾、退伍军人身份或其他任何法律保护的分类而受到不合理的待遇或影响时，就会发生算法歧视。

- 数据隐私保护。系统应该通过内置的保护措施，以免受到滥用数据

行为的影响，而且应该确保用户应该对关于自己数据的使用方式拥有自主权。自动系统的设计者、开发者和部署者应该以适当的方式并在最大程度上获取用户的同意，并且尊重用户关于收集、使用、访问、转移和删除用户数据的决定。

- 告知和解释。自动化系统的设计者、开发者和部署者应该提供可访问的通俗语言文件，包括对整个系统功能和自动化所起作用的清晰描述，关于这些系统使用的告知，负责该系统的个人或组织，应当对结果给出清晰、及时和可访问的解释。

欧洲

欧洲拟议的《人工智能法》主要侧重于加强围绕数据质量、透明度、人类监督和责任的规则。它还旨在解决从医疗和教育到金融和能源等各个领域的道德问题和实施挑战。《人工智能法》的基础是一个分类系统，用来明确人工智能技术可能对个体的健康和安全或基本权利构成的风险水平。该框架包括四个风险等级：不可接受的、高的、有限的和最低的。具有有限和最低风险的人工智能系统，如垃圾邮件过滤器或视频游戏等是允许使用的，除了透明度义务外，没有什么要求。而如政府的社会评分和公共场所的实时生物识别系统会被认为构成不可接受的风险的系统，禁止使用且几乎没有例外。高风险的人工智能系统是被允许的，但开发者和使用者必须遵守规定，要进行严格的测试，对数据质量进行适当的记录，并制定详细的人类监督问责框架。被视为高风险的人工智能包括自动驾驶汽车、医疗设备和关键基础设施机械，以上仅为枚举。拟议的法案还概述了围绕所谓的通用人工智能的规定，这些人工智能系统可用于不同的目的，具有不同程度的风险。此类技术包括，例如，像 ChatGPT 这样的大型语言模型生成型人工智能系统。《人工智能法》提出了严厉的违规处罚措施。对于公司来说，罚款可高达 3000 万欧元或全球收入的 6%。向监管机构提交虚假或

误导性文件也会导致罚款。

欧盟 GDPR 法规

欧盟通用数据保护监管法（GDPR）是为个人的数据在处理和数据流动方面提供保护。该监管法于 2016 年 5 月 24 日生效，并于 2018 年 5 月 25 日开始在欧盟所有成员国都具有约束力并直接适用。GDPR 要求从事个人数据处理的所有人必须遵守其规定，并赋予个人数据正在处理的个人一些重要的权利。参与个人数据处理的自然人和法人，包括公司和政府机构，都被要求按照 GDPR 行事。潜在的不合规行为可能导致高额罚金，并导致法院诉讼和名誉损害的后果。GDPR 适用于在欧盟设立的参与个人数据处理的自然人和法人。但是，对于位于欧盟以外国家的公司、机构和个人，当他们处理欧盟公民或居民的个人数据时，他们须按照 GDPR 开展活动。该法案重点保护的是自然人的“个人数据”，号称史上最严的数据保护法案。根据该法案规定的“市场地原则”，只要数据的收集方、数据的提供方（被收集数据的用户）和数据的处理方（比如第三方数据处理机构）这三方之中，有任何一方是欧盟公民或法人，就将受到该法案管辖。这就是说，任何企业只要是在欧盟市场提供商品或服务，或者收集欧盟公民的个人数据，都将受到这部法律的管辖。对于违法企业的罚金，最高可达 2000 万欧元（约合 1.5 亿元人民币）或者其全球营业额的 4%，以这二者中高者为准。

中国

中国在第三届“一带一路”国际合作高峰论坛提出了《全球人工智能治理倡议》

互联网信息服务算法推荐管理规定

近年来，随着中国《网络安全法》、《数据安全法》、《个人信息保护法》等相关法律的不断完善，人民群众对于个人信息、数据安全以及网

络安全不断重视。党中央在《法治社会建设实施纲要（2020-2025 年）》明确提出制定完善对算法推荐等新技术应用的规范管理办法。2021 年九部委出台《关于加强互联网信息服务算法综合治理的指导意见》，为《互联网信息服务算法推荐管理规定》（以下简称“《规定》”）的制定奠定了良好的法制基础。

中国国家互联网信息办公室、工业和信息化部、公安部、国家市场监督管理总局联合发布《互联网信息服务算法推荐管理规定》（以下简称《规定》），自 2022 年 3 月 1 日起施行。 《规定》明确，应用算法推荐技术，是指利用生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等算法技术向用户提供信息。

《规定》明确了算法推荐服务提供者的信息服务规范，要求算法推荐服务提供者应当建立健全用户注册、信息发布审核、数据安全和个人信息保护、安全事件应急处置等管理制度和技术措施，定期审核、评估、验证算法机制机理、模型、数据和应用结果等；建立健全用于识别违法和不良信息的特征库，发现违法和不良信息的，应当采取相应的处置措施；加强用户模型和用户标签管理，完善记入用户模型的兴趣点规则和用户标签管理规则；加强算法推荐服务版面页面生态管理，建立完善人工干预和用户自主选择机制，在重点环节积极呈现符合主流价值导向的信息；规范开展互联网新闻信息服务，不得生成合成虚假新闻信息或者传播非国家规定范围内的单位发布的新闻信息；不得利用算法实施影响网络舆论、规避监督管理以及垄断和不正当竞争行为。

《规定》要求，具有舆论属性或者社会动员能力的算法推荐服务提供者应当在提供服务之日起十个工作日内通过互联网信息服务算法备案系统填报备案信息，履行备案手续；备案信息发生变更的，应当在规定时间内办理变更手续。算法推荐服务提供者应当依法留存网络日志，配合有关部门

开展安全评估和监督检查工作，并提供必要的技术、数据等支持和协助。

互联网弹窗信息服务管理规定

中国国家互联网信息办公室于 2022 年 9 月 9 日，《互联网弹窗信息推送服务管理规定》（以下简称“规定”）正式发布，已于 2022 年 9 月 30 日正式施行。《规定》总共十条，条文简短，影响重大。《规定》明确了提供信息弹窗的介质包括操作系统、应用软件和网站，这也纠正了部分人的认知误区，《规定》的主要适用主体和监管对象不仅是 PC 端桌面和网站弹窗，还包括移动 APP、小程序、电脑软件等各种应用软件。这对企业来说，尤其要注意，以各种操作系统或硬件类型（例如智能穿戴设备、智慧电视等）为载体的应用软件，通过所有对外接口进行的弹窗信息，均应当符合《规定》的要求。其中，《规定》第三条明确互联网弹窗信息推送服务应当遵守的法律法规明确为“宪法、法律和行政法规”。事实上，随着近两年互联网行业法律法规的不断出台，与互联网弹窗广告、信息内容安全和用户权益保护相关的上位法体系已比较完善，此类调整给监管实务与企业合规治理过程中，提供了更加清晰的合规指向。《规定》第五条第五款：提供互联网弹窗信息推送服务的，应当健全弹窗信息推送内容管理规范，完善信息筛选、编辑、推送等工作流程，配备与服务规模相适应的审核力量，加强弹窗信息内容审核。

美国人工智能监管侧重于人工智能反对歧视欺诈滥用，立法关注应对人工智能带来的危害，敦促企业遵守相关法律法规。

欧洲人工智能监管侧重于可审计可理解，要求于人类交互的人工智能系统需要符合透明度规则，对于情感识别系统或者生物分类系统，系统提供者应将系统的运营情况和结果告知使用者，高风险类型的人工智能在上市前，需要进行风险评估。

中国人工智能监管侧重于明确标准具体指导，明确人工智能算法推荐服务者的主体责任，定期对算法进行审核、评估和验证，并要求算法推荐服务 i 提供者加强算法规则的透明度和解释性。

2.1.2 行业标准

随着人工智能 AI 技术的快速发展，各行业专家希望通过 AI 来加速本行业信息技术与业务能力的建设，但与此同时安全性保障也开始被大家关注起来。本部分主要介绍 AI 安全相关的行业标准。由于标准的起草与发布又很很强的时效性，也必定会不断完善，所以截止本文撰稿期间收集到的标准汇总如下，并在后续章节中针对以下各行业各标准做出介绍。

汽车行业：

- ISO/AWI PAS 8800 Road Vehicles — Safety and artificial intelligence

通信行业：

- YD/T 4044-2022 基于人工智能的知识图谱构建技术要求

医疗行业：

- YD/T 4043-2022 基于人工智能的多中心医疗数据协同分析平台参考架构
- IEEE 2801- 2022 IEEE Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence

金融行业：

- JR/T 0221-2021 人工智能算法金融应用评价规范

2.1.2.1 汽车行业

ISO/AWI PAS 8800 Road Vehicles — Safety and artificial intelligence

2022 年 9 月 20 日，国际标准 ISO/PAS 8800（Road Vehicles — Safety and Artificial Intelligence—道路车辆-人工智能 AI 安全）形成工作组草案并启动国际范围内的公开征集意见。中国作为该标准第 9 小组（AI 运行监控和持续安全保障）牵头方深度参与研究制定工作。

ISO/PAS 8800 是 ISO/TC22/SC32/WG14 道路车辆人工智能 AI 安全工作组的首个国际标准，自动驾驶领域核心的 AI 安全技术受到国际广泛关注，来自中国、美国、德国、英国、奥地利、日本、韩国等 17 个国家的专家参与起草工作。由中汽中心、一汽、华为、商汤、地平线组成的中国专家代表团作为该标准第 9 小组（AI 运行监控和持续安全保障）牵头方参与研究制定工作。

Subteam	Name
Subteam 01	Concepts and Definitions
Subteam 02	Scoping
Subteam 03	Functional Safety and SOTIF
Subteam 04	Safety Lifecycle
Subteam 05	Definition of safety-related properties of AI functions
Subteam 06	Selection of AI techniques and design related considerations
Subteam 07	Data-related properties

Subteam 08	Evaluating the performance/V&V
Subteam 09	Measures during operation and continuous assurance

表 1 国际标准内容

该标准融合了功能安全 ISO 26262、预期功能安全(SOTIF)ISO 21448、ISO/IEC TR 5469 人工智能 AI 功能安全、ISO TS 5083 自动驾驶系统安全，对车辆 AI 系统提出了安全要求，涵盖 AI 安全属性定义、AI 技术选择和设计、数据定义和挑选、AI 安全措施、AI 安全性能评价、AI 验证和确认 V&V、运行监控和 AI 安全保障等关键内容，以避免不合理的风险。

该标准定义了影响道路车辆环境中人工智能（AI）性能不足和故障行为的安全相关特性和风险因素。描述了一个处理开发和部署生命周期所有阶段的框架。这包括对功能的适当安全要求，与数据质量和完整性相关的考虑因素，控制和缓解故障的架构措施，用于支持人工智能的工具，验证和验证技术，以及支持系统整体安全要求的依据。

中国牵头负责的第 9 组-AI 运行监控和持续安全保障，重点关注基于目标市场的 AI 安全运行监控措施定义、车端和云端量化安全监控机制开发、现场数据搜集、AI 重训和技术完善等内容。

2.1.2.2 通信行业

YD/T 4044-2022 基于人工智能的知识图谱构建技术要求

该标准规定了基于人工智能的知识图谱系统构建的技术要求、基本功能要求、非功能要求，用于规范基于人工智能的知识图谱的框架构建流程。 可以为科技企业、用户机构、第三方机构等提供指导，包括对基于人工智能的知识图谱系统

进行设计、开发、测试等内容。

在该标准的 8.6 章节 安全性要求部分对基于人工智能的知识图谱系统的安全性提出要求，主要关注在外部安全标准参考（例如 GB/T 22239—2019 和 GB/T 35273—2020），敏感数据传输相关的访问控制要求，数据传输实体限制要求，数据机密性、完整性、可用性要求，数据权限要求等。

2.1.2.3 医疗行业

YD/T 4043-2022 基于人工智能的多中心医疗数据协同分析平台参考架构

该标准件规定了构建多中心医学数据协同分析平台的参考构架，包括系统架构、数据存储、数据标准化、数据分析原则、平台特性、功能需求和安全性等。主要为多中心医疗数据协同分析平台的软件、硬件和网络构架提供指导。

该标准安全部分要求主要包含两部分内容，在 4.2 章节数据隔离性部分提出各数据落地要求、数据服务器单点避免要求、以及医疗标准参考要求。并在 4.10 章节安全性部分提出数据高安全性要求、保证原始数据与业务数据安全要求，以及系统安全标准符合要求。

2.1.2.4 金融行业

JR/T 0221-2021 人工智能算法金融应用评价规范

2021 年 3 月 26 日，中国人民银行正式发布金融行业标准（JR/T 0221—2021）《人工智能算法金融应用评价规范》。

该标准从安全性、可解释性、精准性和性能等方面建立了人工智能算法金融应用评价框架，明确了智能算法应用的基本要求、评价方法、判定准则，为金融机构加强智能算法应用风险管理提供指引。

该标准的发布旨在引导金融机构加强对人工智能算法金融应用的规范管理

和风险防范，加快金融数字化转型步伐，持续推动金融服务更为贴心、更加智慧、更有温度，打造数字经济时代金融创新发展新引擎。

该标准由全国金融标准化技术委员会归口管理，由中国人民银行科技司提出并负责起草，行业内有关单位共同参与。

该标准的第 6 章节 安全性评价部分针对目标函数、常见攻击范围、算法依赖库、算法可追溯性、算法内控等内容提供了基本要求、评价方法与判定准则的指导。

2.1.2.5 国内标准

AI 安全基础标准：

1)我国首个人工智能安全国家标准：**信安标委 TC260**《信息安全技术 机器学习算法安全评估规范》（报批稿），规定了机器学习算法技术在生存周期各阶段的安全要求，以及应用机器学习算法技术提供服务时的安全要求，并给出了对应评估方法。《信息安全技术 人工智能计算平台安全框架》国家标准（征求意见稿），规范了人工智能计算平台安全功能、安全机制、安全模块以及服务接口，指导人工智能计算平台设计与实现。《信息安全技术 生成式人工智能预训练和优化训练数据安全规范》在研阶段。

2)信标委 TC28/SC42，《人工智能 管理体系》在研阶段，《人工智能 深度学习框架多硬件平台适配技术规范》在研阶段。

3)中国电子工业标准化技术协会（CESA）《信息技术 人工智能 风险管理能力评估》已发布。

4)中国人工智能产业发展联盟（AIIA）《可信人工智能 组织治理能力成熟度模型》在研阶段，《大规模预训练模型技术和应用评估方法 第 5 部分：安全

可信》在研阶段。

负责/归口	标准类型	标准编号	标准名称	阶段
全国信安标委（TC260）	国家标准	20211000-T-469	信息安全技术 机器学习算法安全评估规范	报批稿
全国信安标委（TC260）	国家标准	20230249-T-469	信息安全技术 人工智能计算平台安全框架	征求意见稿
全国信安标委（TC260）	国家标准	--	信息安全技术 生成式人工智能预训练和优化训练数据安全规范	在研
全国信标委人工智能分委会（TC28/SC42）	国家标准	20221791-T-469	人工智能 管理体系	在研
全国信标委人工智能分委	国家标准	20221795-T-469	人工智能 深度学习框架多硬件平台适配技术规范	在研

会（TC28/SC42）				
中国电子工业标准化技术协会（CESA）	团体标准	T/CESA 1193-2022	信息技术 人工智能 风险管理能力评估	已发布
中国人工智能产业发展联盟（AIIA）		--	可信人工智能 组织治理能力成熟度模型	在研
中国人工智能产业发展联盟（AIIA）		--	大规模预训练模型技术和应用评估方法 第 5 部分：安全可信	在研

表 2 AI 安全基础标准

AI 应用相关安全标准：

1)全国信安标委（TC 260）在生物特征识别方向，发布了 GB/T 40660—2021《信息安全技术 生物特征识别信息保护基本要求》，以及人脸、声纹、基因、步态等 4 项数据安全国家标准。在智能汽车方向，发布了国家标准 GB/T 41871—2022《信息安全技术 汽车数据处理安全要求》，有效支撑《汽车数据安全管理若干规定（试行）》，提升了智能汽车相关企业的数据安全水平。

2)全国信标委生物特征识别分委会（TC 28/SC37）发布了《信息技术 生物特征识别呈现攻击检测》、《信息技术 生物特征识别 人脸识别系统应用要求》等

标准。

3)中国通信标准化协会 CCSA 在生物识别、人工智 能终端、人工智能服务平台、数据安全保护等领域开展了数据安全相关标准化工作。在人工智能终端领域，开展《人工智能终端产品 个人信息保护要求和评估方法》与《人工智能终端设备安全环境技术要求》标准研制，对人工智能终端的个人信息保护与终端设备环境的安全能力提出要求。在人工智能服务平台领域，开展《人工智能服务平台数据安全要求》标准研制，对人工智能服务端的数据安全管理与评估提出要求。

4)上海市市监局《人工智能数据通用安全要求》（征求意见稿）、《人脸识别分级分类应用标准》（草案）。

5)中国电子工业标准化技术协会（CESA）《信息安全技术 人脸比对模型安全技术规范》已发布。

6)新一代人工智能产业技术创新战略联盟（AITISA）发布了《人工智能视觉隐私保护 第 1 部分：通用技术要求》、《生物特征识别服务中的隐私保护技术指南》、《生物特征模板的安全使用要求》等标准，《信息技术 数字视网膜系统 第 11 部分：安全与隐私保护》在草案阶段。

负责/归口	标准类型	标准编号	标准名称	阶段
全国信安标委（TC260）	国家标准	GB/T 38542-2020	信息安全技术 基于生物特征识别的移动智能终端身份鉴别技术框架	发布
全国信安标委（TC260）	国家标准	GB/T 38671-2020	信息安全技术 远程人脸识别系统技术要求	发布
全国信安标委（TC260）	国家标准	GB/T 40660-2021	信息安全技术 生物特征识别信息保护基本要求	发布
全国信安标委（TC260）	国家标准	GB/T 41819-2022	信息安全技术 人脸识别数据安全要求	发布
全国信安标委（TC260）	国家标准	GB/T 41807-2022	信息安全技术 声纹识别数据安全要求	发布
全国信安标委（TC260）	国家标准	GB/T 41806-2022	信息安全技术 基因识别数据安全要求	发布

全国信安标委（TC260）	国家标准	GB/T 41773-2022	信息安全技术 步态识别数据安全要求	发布
全国信安标委（TC260）	国家标准	GB/T 41871-2022	信息安全技术 汽车数据处理安全要求	发布
全国信安标委（TC260）	国家标准	20230253-T-469	信息安全技术 基于个人信息的自动化决策安全要求	在研
全国信标委生物特征识别分委会(TC28/SC37)	国家标准	GB/T 41815.1-2022	信息技术 生物特征识别呈现攻击检测 第1部分：框架	发布
全国信标委生物特征识别分委会(TC28/SC37)	国家标准	GB/T 41815.2-202	信息技术 生物特征识别呈现攻击检测 第2部分：数据格式	发布
全国信标委生物特征识别分委会	国家标准	GB/T 41815.3-2023	信息技术 生物特征识别呈现攻击检测 第3部分：测试与报	发布

(TC 28/SC37)			告	
全国信标委生物特征识别分委会（TC 28/SC37）	国家标准	GB/T 37036.3-2019	信息技术 移动设备生物特征识别 第3部分：人脸	发布
全国信标委生物特征识别分委会（TC 28/SC37）	国家标准	GB/T 37036.8-2022	信息技术 移动设备生物特征识别 第8部分：呈现攻击检测	发布
全国信标委生物特征识别分委会（TC 28/SC37）	国家标准	GB/T 5271.37-2021	信息技术 词汇 第37部分：生物特征识别	发布
全国信标委生物特征识别分委会（TC 28/SC37）	国家标准	20221220-T-469	信息技术 生物特征识别 人脸识别系统应用要求	在研

中国通信标准化协会（CCSA）	行业标准	YD/T 4087-2022	移动智能终端人脸识别安全技术要求及测试评估方法	发布
中国通信标准化协会（CCSA）	行业标准	2023-0041TYD	人工智能开发平台通用能力要求 第 2 部分：安全要求	在研
中国通信标准化协会（CCSA）	行业标准	2023-0039TYD	面向人脸识别系统的人脸信息保护基础能力要求	在研
中国通信标准化协会（CCSA）	行业标准	--	人脸识别线下支付安全要求	草案
中国通信标准化协会（CCSA）	行业标准	2021-0630TYD	电信网和互联网人脸识别数据安全检测要求	在研
上海市市监局	地方标准	--	人工智能数据通用安全要求	征求意见稿
上海市市监局	地方标准	--	人脸识别分级分类应用标准	草案
中国电子工业标准化技	团体标准	T/CESA 1124-2020	信息安全技术 人脸比对模型安全技术规范	发布

术协会（CESA）				
新一代人工智能产业技术创新战略联盟（AITISA）	团体标准	T/AI 110.1-2020	人工智能视觉隐私保护 第 1 部分：通用技术要求	发布
新一代人工智能产业技术创新战略联盟（AITISA）	团体标准	T/AI 110.2-2022	人工智能视觉隐私保护 第 2 部分： 技术应用指南	发布
新一代人工智能产业技术创新战略联盟（AITISA）	团体标准	T/AI 113- 2021	生物特征识别服务中的隐私保护技术指南	发布
新一代人工智能产业技术创新战略联盟（AITISA）	团体标准	T/AI 111-2020	生物特征模板的安全使用要求	发布

新一代人工智能产业技术创新战略联盟（AITISA）	团体标准	2023011205	信息技术 数字视网膜系统 第 11 部分：安全与隐私保护	草案

表 3 AI 应用安全相关标准

国际标准：

国际标准组织（ISO）在人工智能领域已开展大量标准化工作，并专门成立了 ISO/IEC JTC1 SC42 人工智能分技术委员会。目前，与人工智能安全相关的国际标准及文件主要为基础概念与技术框架类通用标准，在内容上集中在**人工智能管理、可信性、安全与隐私保护**三个方面。

1）在人工智能管理方面，国际标准主要研究人工智能数据的治理、人工智能系统全生命周期管理、人工智能安全风险管理等，并对相应的方面提出建议，相关标准包括 ISO/IEC 38507:2022《信息技术治理 组织使用人工智能的治理影响》、ISO/IEC 23894:2023《人工智能 风险管理》等。

2）在可信性方面，国际标准主要关注人工智能的透明度、可解释性、健壮性与可控性等方面，指出人工智能系统的技术脆弱性因素及部分缓解措施，相关标准包括 ISO/IEC TR 24028:2020《人工智能 人工智能中可信赖性概述》等。

3）在安全与隐私保护方面，国际标准主要聚焦于人工智能的系统安全、功能安全、隐私保护等问题，帮助相关组织更好地识别并缓解人工智能系统中的安全威胁，相关标准包括 ISO/IEC 27090《人工智能 解决人工智能系统中安全威胁和故障的指南》、ISO/IEC TR 5469《人工智能 功能安全与人工智能系统》、ISO/IEC 27091《人工智能 隐私保护》等。

欧洲

欧洲电信标准化协会（ETSI）近期关注的重点议题包括人工智能数据安全、完整性和隐私性、透明性、可解释性、伦理与滥用、偏见缓解等方面，已发布多份人工智能安全研究报告，包括 ETSI GR SAI 004《人工智能安全：问题陈述》、ETSI GR SAI 005《人工智能安全：缓解策略报告》等，描述了以人工智能为基础的系统安全问题挑战，并提出了一系列缓解措施与指南。

欧洲标准化委员会（CEN）、欧洲电工标准化委员会（CENELEC）成立了新的 CEN-CENELEC 联合技术委员会 JTC 21“人工智能”，并在人工智能的风险管理、透明性、健壮性、安全性等多个方面提出了标准需求。

美国

美国国家标准与技术研究院（NIST）关注人工智能安全的可信任、可解释等问题。最新的标准项目有：NIST SP1270《建立识别和管理人工智能偏差的标准》，提出了用于识别和管理人工智能偏见的技术指南；NIST IR-8312《可解释人工智能的四大原则》草案，提出了可解释人工智能的四项原则；NIST IR-8332《信任和人工智能》草案，研究了人工智能应用安全风险与用户对人工智能的信任之间的关系；NIST AI 100-1《人工智能风险管理框架》，旨在为人工智能系统设计、开发、部署和使用提供指南。

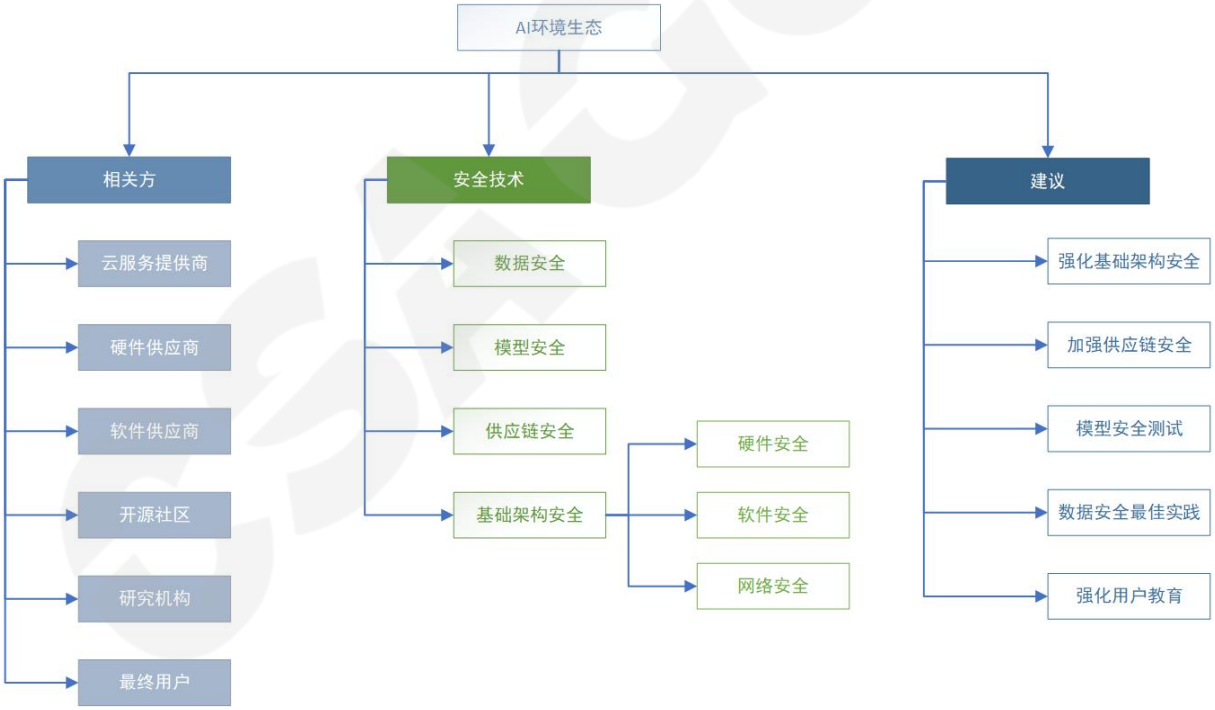
2.1.3 国际共识

2017 年 1 月，在加利福尼亚州阿西洛马举行的 Beneficial AI 会议上，近千名人工智能和机器人领域的专家,联合签署了阿西洛马人工智能 23 条原则,呼吁全世界在发展人工智能的同时严格遵守这些原则,共同保障人类未来的伦理、利益和安全。阿西洛马人工智能原则(Asilomar AI Principles)是著名的阿西莫夫的机器人三大法则的扩展版本。

原则目前共 23 项，分为三大类，分别为：科研问题（Research Issues）、伦理和价值（Ethics and values）、长期问题（Longer-term Issues）。科研问题主要包括研究问题、研究经费、科学与政策的联系、科研文化和避免战争；伦理与价值主要包括安全性、故障透明性、司法透明性、责任、价值归属、人类价值观、个人隐私、自由和隐私、分享利益、共同繁荣、人类控制、非颠覆和 AI 军备竞赛；长期问题包括能力警惕、重要性、风险、递归的自我提升和公共利益。其中与安全相关的包括第六条安全性、第七条故障透明性、第十二条个人隐私、第十三条自由和隐私、第二十一条风险、第二十二条递归的自我提升。

2.2 AI 安全的技术生态

2.2.1 AI 环境的安全



2.2.1.1 AI 环境安全的分类

作为 AI 安全的技术生态的基础，AI 环境生态的安全是整个 AI 安全的实施基础。概览而言，我们将 AI 环境生态分为 2 个大的方面：

1）AI 环境的干系人（相关方）

AI 环境的干系人，主要有如下的几类：

A）政府机构：负责制定相关政策、法规和，确保 AI 环境的安全和可持续发展；

B）AI 开发和研究者：遵守道德准则和伦理规范，确保 AI 系统的安全性和可控性。

C）用户和企业：加强对 AI 系统的安全意识和风险认知，采取相应的安全措施来保护自身和企业的利益。

在本文中，我们将重点关注 AI 的相关方的安全生态。

2）AI 环境的安全技术支持

AI 环境中的安全技术主要包括：

A）数据安全

数据安全是确保 AI 系统所使用的数据的保密性、完整性和可用性。这包括对数据进行加密、访问控制、备份和恢复等措施，以防止未经授权的访问、数据泄露或数据损坏。

B）算法安全

算法安全关注的是 AI 系统所使用的算法的安全性。这包括对算法进行安全审计、漏洞分析和修复，以确保算法的正确性、鲁棒性和防御性，防止算法被恶意攻击者利用或操纵。

C）模型安全

模型安全是确保 AI 系统所使用的模型的安全性和可信度。这包括对模型进

行鲁棒性测试、模型解释和可解释性分析，以及对模型进行防御性设计，以减少模型受到对抗性攻击的风险。

D）应用安全

应用安全关注的是 AI 系统的应用环境的安全性。这包括对 AI 系统的部署和运行环境进行安全配置、访问控制和监控，以防止未经授权的访问、恶意攻击或滥用。

E）供应链安全

供应链安全是确保 AI 系统的整个供应链过程中的安全性。这包括对 AI 系统或组件的开发、集成和部署过程进行安全审计和监控，以防止恶意代码、恶意组件或恶意行为被引入到 AI 系统中。

F）基础架构安全

基础架构安全关注的是 AI 系统所依赖的基础设施的安全性。这包括对网络、服务器、存储和通信等基础设施进行安全配置、漏洞管理和监控，以保护 AI 系统的运行环境免受攻击和威胁。可以把基础架构安全简化为：包括硬件、软件安全、网络安全三个部分。

以上这些安全技术共同构成了保障 AI 环境安全的重要组成部分，通过综合应用这些技术，可以提高 AI 系统的安全性、可靠性和可信度，减少潜在的风险和威胁。

2.2.1.2 AI 环境安全的相关方生态

AI 环境的相关方，除了政府机构、AI 开发者和研究者、AI 用户和企业之外，也可以将其从环境提供和使用的角度，分为：云服务提供商、硬件供应商、软件供应商、开源社区、研究机构、最终用户。

2.2.1.2.1 云服务提供商的安全生态

在人工智能（AI）的发展与应用日益深入的当下，安全问题凸显为一个无法回避的关键议题。云服务提供商在推动 AI 技术的创新与应用的同时，必须提供一套多维度的安全防护体系，包括数据安全、网络安全、计算安全等，保障 AI 应用在云平台的全生命周期安全。

从工具和应用角度上来看，云服务提供商应整合丰富的安全工具与产品，如防火墙等网络安全工具、加密及掩码等数据安全工具以及云原生的身份及访问管理（IAM）等，满足客户多样化的安全需求。

此外，云服务提供商需要为 AI 应用提供自动化开发和运营能力，包括：

- 开发并提供自动化的合规性检查工具，以确保 AI 项目符合行业标准与法规要求；
- 以及实施实时的合规性监控，确保 AI 应用在运行阶段的持续合规；
- 监控云环境中的安全事件，并在检测到异常行为时触发报警；
- 依据预设的安全策略，自动执行相应的安全响应措施，例如隔离受影响的系统、阻断非法访问等。

云服务提供商在推进 AI 技术的广泛应用的同时，必须重视 AI 安全问题。针对不同行业的特定需求，为 AI 客户提供定制化的安全解决方案，以满足各行业在数据保护、合规性等方面的特殊要求。与各类安全技术提供商建立合作关系，共同打造丰富的安全产品与服务体系，满足客户多样化的安全需求。建立安全社区，加强与全球安全研究者、机构的交流与合作，共同推动云计算与 AI 安全领域的技术创新与进步。

2.2.1.2.2 硬件供应商的安全生态

在构建 AI 安全环境中，硬件供应商的角色不仅限于提供传统的基于硬件的网络和数据安全工具。他们还应增强以下关键安全能力：

运行环境安全：除了关注安全启动外，硬件的运行环境安全还涉及到运行时的安全性。安全启动包括检验固件和启动配置的完整性和真实性，防止启动过程中的篡改和攻击；及硬件元件和连接接口的防篡改设计，以抵御物理攻击和非法访问。而运行时安全则包括确保硬件资源的访问控制，阻止未授权的数据访问和操作；以及监测硬件运行环境的温度、湿度等因素，及时调整或警告，以防止环境因素导致的设备故障。

计算安全：对于 AI 的安全性，硬件供应商所提供的安全计算和模型保护能力尤为关键。这包括支持数据在加密状态下的计算，以防止计算过程中的数据泄露；提供硬件加速方案，支持大规模安全计算的高效执行；在存储和调用模型时实施加密，防止模型的权重和结构被窃取或篡改；以及通过硬件的手段实现不同计算任务的物理隔离，以防止交叉攻击。

数据安全：从数据的生成到其存储、传输和使用，硬件供应商都应确保整个过程的加密保护，以确保数据的安全。同时，供应商应提供安全的存储方案，例如使用硬件安全模块（HSM）来存储和保护密钥等敏感信息。侧信道攻击利用硬件执行时间差异、功率消耗、电磁泄露等物理信息攻击密码系统，成为数据安全的一大风险。因此，硬件供应商应采用抗侧信道攻击的加密算法和实施物理层面的保护措施，例如电源噪声抑制。此外，物理防护的硬件存储单元和硬件级别的冗余备份机制可以确保数据的持久性和可靠性。

2.2.1.2.3 软件供应商的安全生态

在 AI 安全的生态中，软件供应商扮演着至关重要的角色。除了基本的安全操作系统等运行环境，以及对数据加密、隐私增强等安全技术的支持外，软件供应商需要提供一系列综合性的解决方案来确保 AI 系统的安全运行。

安全开发生命周期：“安全左移”的理念已经成为业界共识。对于 AI 产品而言，同样需要将安全深入融入到产品的开发生命周期中。软件供应商应从源头上减少安全风险，通过强化安全编码规范和实施安全编码培训，确保开发团队具备充足的安全意识和能力。定期并在重大迭代前进行代码审计，以便及早发现并修复潜在的安全问题，确保代码在架构和实现层面的安全性。整合自动化安全测试工具和流程，持续监控代码库的安全状况，及早发现并处理安全漏洞。在关键版本发布前进行渗透测试，确保产品具备足够的安全防护能力，能够抵御实际攻击。

安全运营和维护：软件供应商需要确保其提供的软件或产品的持续的安全性，并能够及时响应各种安全事件和需求。构建实时的安全监控体系，对系统运行状态进行 24/7 监控，及时发现并处理安全事件。引入智能的异常分析工具，及时发现非正常行为，并通过深度分析排查潜在安全威胁。此外，软件供应商亦需要建立安全漏洞的快速响应机制，及时发布和推送安全补丁，降低安全风险的影响；以及实施严格的版本控制和回滚机制，确保在发生安全事件时，能够迅速恢复服务。

安全赋能：为了构建一个更加安全的 AI 生态环境，软件供应商需要将其在安全方面的专业知识和能力拓展到 AI 应用开发者社群中，从而强化整个生态的安全防护能力。包括：进行开发者教育、提供安全工具和安全 SDK、构建社区支持等。如此，软件供应商不仅保证了自身产品的安全，还促进了整个 AI 生态的安全发展，降低了整个生态面临的安全风险。

2.2.1.2.4 开源社区的安全生态

开源社区在AI领域的贡献显著，并且大量的AI项目和框架都源于开源社区。然而，安全问题也在这些社区中呈现出一些独特的挑战。由于开源项目通常依赖于社区贡献者的输入，代码的质量和安全性可能会受到贡献者经验和知识的影响。而代码和设计公开，恶意攻击者则更容易找到和利用安全漏洞。此外，一些项目

可能由于缺乏维护而存在安全风险。

因此在开源社区实践安全，为 AI 的安全发展具有重要意义。首先，应遵循和推广最佳的安全实践和标准，确保开源项目的安全性。再者，实施代码审计，并使用自动化安全扫描工具，以提早发现潜在的安全问题。而后，建立一个清晰的安全漏洞报告和处理流程，以及一个奖励系统来激励找到漏洞的独立安全研究者。最后建立和推广一种安全文化，强调安全的重要性，并鼓励社区成员参与到安全实践中来。

2.2.1.2.5 研究机构的安全生态

研究机构在 AI 领域深入探讨理论和实践，而其安全生态同样显得至关重要。这些机构一方面需要构建安全的科研环境确保自身的研究方法、过程、成果得以受到保护，相对于应用层面，更需要加强科研人员在研究过程中的合规、伦理道德教育，以及安全意识。包括：设立伦理委员会，确保研究活动的合法性和道德性；定期进行合规性检查，确保研究活动符合相关法规和标准；在数据收集、使用和共享中遵循数据伦理原则，保护数据主体的权益；为研究人员提供定期的安全培训和教育，提高其安全意识和能力，鼓励和推广安全的研究实践，建设积极的安全文化等。

2.2.1.2.6 最终用户的安全生态

最终用户是 AI 生态中的关键参与者，他们的安全直接关系到整个生态的稳定和健康。除了技术上为终端用户强化安全能力以及数据隐私保护，为最终用户提供安全的用户体验，包括安全提示、默认的安全设置以及安全自动更新能力也至关重要。最终用户应了解其数据使用方式，并能够自主管理、删除或迁移自己的数据。此外，教育用户识别钓鱼邮件、应用、网站等欺诈行为，增强他们的防范意识也是关键之一。最后，在用户遭受攻击后，各级供应商为用户提供专门的安全支持通道，帮助他们迅速恢复。为了有效地保护最终用户，除了技术手段，

还需要与用户进行持续的交流和教育，让他们具备基本的安全意识和能力。同时，技术提供商和服务商也应该始终以用户的安全和隐私为中心，持续地改进和优化其产品和服务。

2.2.1.3 AI 环境安全的提升建议

综合上述提到的 AI 环境的安全生态。对于 AI 环境安全的工作，有如下一些建议：

1） 加强合作与信息共享

政府机构、学术界和产业界应加强合作，共享关于 AI 安全的最佳实践、威胁情报和技术研究成果。

2）建立 AI 安全评估机制

AI 的开发者、研究者、以及相应的 AI 企业应制定统一的 AI 安全评估标准和流程，对 AI 系统进行安全性评估和认证，确保其符合安全要求。

3）培养 AI 安全专业人才

政府机构和产业界都需要加强对 AI 安全领域的人才培养，培养专业的 AI 安全工程师和研究人员，提高 AI 系统的安全性。

4）定期更新和维护 AI 系统

政府和企业对于 AI 系统应及时修复漏洞和更新安全补丁，保持 AI 系统的安全性和稳定性。

5）建立应急响应机制

具体实操方面，政府机构既要建立起社会层面的应急响应机制，积极应对 AI 系统可能带来的社会面信息失真或管控失效的风险，同时也要指导企业尽快建立

起快速响应的机制，要求企业对于内部尤其是外部的 AI 应用系统应建立快速响应和处置机制，应对 AI 安全事件和威胁，减少损失和风险。

2.2.2 AI 数据的安全

2.2.2.1 AI 数据的安全威胁

数据安全是 AI 数据安全的关键。数据的安全和质量影响着 AI 算法/模型的准确性，在机器学习的每个步骤中，都存在着隐私泄露的风险：

1. 数据收集阶段：在数据收集阶段，存在着身份证号、手机号等敏感个人信息数据被收集，会造成用户隐私泄露；此外，通过 AI 分析关联大规模收集到的数据，易造成额外隐私泄露。另外，由于数据收集、传输途径的不可靠，攻击者可能会注入一些恶意数据，导致数据投毒等攻击。

2. 模型训练和测试阶段：一方面，训练数据、计算资源等可能分属于不同参与方，集中式训练会导致训练数据的泄露；另一方面，模型的中间参数被恶意获取，导致反推出训练数据。在这个阶段面临的主要攻击是模型反演攻击。

3. 预测阶段：预测模型一般是存储在云服务器中，用户希望计算自己的私有数据得到预测结果，存在着模型泄露与用户私有数据泄露的风险；此外，终端用户可以构造反向推理攻击获取训练数据或模型数据，主要的攻击包括模型成员推断攻击。

机器学习阶段	可能面临的攻击/安全风险
数据收集/预处理阶段	数据投毒攻击、个人敏感信息泄露
训练阶段	模型逆向攻击、模型提取攻击
预测阶段	模型逆向攻击、模型提取攻击、成员推断攻击

表 4 机器学习三个阶段

- 数据投毒攻击是指攻击者修改一定数量的数据，使得模型训练出错。
- 模型逆向攻击是指攻击者从模型预测结果中提取和训练数据有关的信息。
- 模型提取攻击是指攻击者获得对某个目标模型的黑盒访问权后，取得模型内部的参数或结构，或是试图构造出一个与目标模型近似甚至完全等价的机器学习模型。
- 成员推断攻击是指攻击者通过访问模型预测 API，从预测结果中获知某个特征数据是否包含在模型的训练集中。

2.2.2.2 AI 数据的安全保护

面对上述 AI 数据的安全威胁，本小节主要梳理介绍面向 AI 数据的安全保护技术，而对于其他与数据相关，但会给 AI 算法、模型等带来威胁的攻击手段（如数据投毒、对抗样本等）的防护技术将在其他相应章节中给出深入分析。

（1）联邦学习

联邦学习是一种特殊的分布式机器学习框架，它仅通过交互模型中间参数（或加密参数）来完成在多方联合训练，可以保障每方的原始数据不出本地。一般是由多个客户端和一个中央服务器组成，各个客户端从中央服务器下载现有的训练模型，利用自己的本地数据和计算资源对模型进行训练，并将训练后的模型

参数加密上传至中央服务器，经中央服务器聚合后产生新的模型参数。通过重复这个过程直至停止训练。

按照训练数据在不同数据方之间的特征空间和样本空间分布情况，将联邦学习分为横向联邦 DC 学习、纵向联邦学习以及迁移联邦学习。

- 横向联邦学习：本质上是样本的联合，适用于不同数据集之间，特征重合较多而样本重合较少的情形。目前横向联邦学习的典型框架是 FedAvg，典型应用包括面向手机/物联网设备的下一单词预测、人脸识别等，面向组织的多家银行联合风控信贷等。

- 纵向联邦学习：本质上是特征的联合，适用于不同数据集之间，样本重合较多而特征重合较少的情形，主要适用于面向组织的场景，如融合银行、保险公司、政府等多方数据的营销风控场景。

- 迁移联邦学习：适用于不同数据集之间，样本和特征重合均较少的情形。如不同地区的银行和电商间的联合建模。

开源框架

目前业界中主要的联邦学习开源框架有 FATE、PySyft、TensorFlow Federated、PaddleFL 等。

- FATE 是微众银行在 2019 年开源的一种工业级联邦学习框架，全面支持横向联邦学习、纵向联邦学习及迁移联邦学习，涵盖了 LR、GBDT、CNN 等常见机器学习算法，覆盖常规商业应用建模场景需求。此外，FATE 提供一站式联邦模型解决方案，包括特征工程、离线训练、在线预测等模块。

- PySyft 是由 OpenMined 提出的一种基于 python 的隐私保护深度学习框架，涵盖了包括差分隐私、同态加密和多方安全计算等多种隐私保护机制。目前发布的版本仅支持横向联邦学习，支持 LR、DNN 等算法。

- TFF（TensorFlow Federated）是于2019年由谷歌发布的基于Tensorflow首个大规模移动设备端联邦学习系统，旨在促进联邦学习的开放性研究和实验，其受众主要是研究人员。联邦学习类型方面，目前只支持横向联邦学习；模型方面，提供了FedAvg, Fed-SGD等聚合算法，同时也支持神经网络和线性模型。在计算范式方面， TFF 支持单机模拟和移动设备训练，不支持基于拓扑结构的分布式训练； 在隐私保护机制方面， TFF 采用差分隐私以保证数据安全。

- PaddleFL是由百度发布的联邦学习开源框架，由于其底层编程模型采用的是飞桨训练框架，结合飞桨的参数服务器功能，其可以实现在Kubernetes 集群中联邦学习系统的部署。支持横向联邦和纵向联邦，其中横向联邦学习主要涵盖了FedAvg、DP-SGD等聚合算法，纵向联邦方面主要包括基于PrivC的逻辑回归和基于ABY3协议的神经网络。

除了上述的几款开源框架外，还有美国南加州大学等联合研发的FedML，英国牛津大学研发Flower，字节跳动研发的Fedlearner，阿里巴巴达摩院研发的FederatedScope等开源框架。

厂商

联邦学习主要集中在互联网龙头企业、初创公司以及运营商、金融科技等行业数据高度聚合的企业。从2018年起，腾讯、阿里、蚂蚁、京东、百度、字节跳动等互联网龙头企业，富数科技、同盾等初创型科技企业，开展联邦学习技术的战略和应用，推动相关行业解决方案和项目。此外，中国移动、微众银行、工商银行、农业银行、建设银行、招商银行、平安集团等行业数据高度聚合企业利用联邦学习技术开展数据增值业务。

- 腾讯研发的AngelFL联邦学习平台是构建在Angel智能学习平台的基础之上，是一种“无可信第三方”的联邦学 习框架。整个系统以 Angel 的高维

稀疏训练平台作为底层，抽象出“算法协议”层，供实现各种常见机器学习算法，支持逻辑回归、GBDT 等算法。

- 京东数字科技集团自主研发的 Fedlearn 联邦学习平台，融合了密码学、机器学习、区块链等联邦学习算法，搭建出一套安全、智能、高效的链接平台，在各机构数据不用向外传输的前提下，通过联合多方机构数据，实现共同构建模型等多方数据联合使用场景，获得加成效应。相较于传统的数据共享交换方法，Fedlearn 平台创新性地提出了并行 加密算法、异步计算框架、创新联邦学习等技术架构，在保证数据安全的前提下提升学习效率，并逐步达到融合亿级规模数据的能力。支持 FedAvg、DNN、线性模型、逻辑回归及随机森林等算法。

- 平安科技研发的蜂巢联邦智能平台，是数据安全保护、企业数据孤岛、数据垄断、数据壁垒等问题的商用级解决方案。蜂巢充分支持了国密 SM2、国密 SM4 以及混淆电路、差分隐私和同态加密等不同的加密方式,以满足企业各个业务场景的不同需求，此外，采用 GPU 等异构计算芯片来加速联邦学习的加密和通信过程,从而达到效率升级的效果。支持横向联邦建模和纵向联邦建模两种模式。

- 光之树科技研发的云间联邦学习平台，该平台是基于机器学习、深度学习算法和加密协议的安全计算框架。拥有自动建模的功能，支持多种机器学习和深度学习训练和模型部署。应用于普惠金融、贸易金融、保险反欺诈、供应链金融等场景，支持横向和纵向联邦学习。

（2）多方安全计算

多方安全计算起源于图灵奖获得者姚期智院士在 1982 年提出的百万富翁问题，是指在无可信第三方情况下，多个参与方共同计算一个目标函数，并且保证每一方仅获取自己的计算结果，对其他参与方的结果/输入一无所知。因此，在 AI 数据保护方面，通常多方安全计算协议被重点应用在高效并行分布式机器学习中。一方面，在学习过程中保障了（多方）训练数据的安全性/隐私性；另一

方面在安全推理中，实现对模型和预测数据的保护，但是多方安全计算不能防止对结果模型的推理攻击。

多方安全计算是由混淆电路、不经意传输协议、秘密共享等多种密码学基础工具综合应用而来。基于多方安全计算对 AI 数据进行保护，按照技术路线的不同，可以分为以下三大类：

1. 基于秘密共享的方法：各方通过秘密共享方案将他们的数据分成多份分发给其他参与方，各方利用这些秘密份额进行本地计算，并最终重构出结果。这类方法通常具有较低的计算复杂度，其通信量和通信轮次与电路深度成正比，适用于大规模数据。对于线性运算效率较高，涉及到非线性运算时效率较低。适用于机器学习的安全训练、安全预测两个阶段。典型协议有 SPDZ、SecureNN、Falcon 等。

2. 基于混淆电路的方法：通过混淆电路构建加密的神经网络或其他机器学习模型，在不解密的条件下进行训练或预测任务。这类方法大多用于两方场景，具有常数轮通信次数，但是通信开销与电路大小成正比，通信量较大，此外由于混淆电路需逐比特计算，计算线性操作计算复杂度较高，但对于非线性函数如比较时较为高效。通常用于安全预测，只适用于简单的机器学习模型训练如逻辑回归。典型的协议有 FairplayMP 和典型框架 DeepSecure 等。

3. 基于混合协议的方法：还有一些协议充分利用不同技术的优点，将上述两种技术及同态加密等其他技术结合起来，应用于机器学习模型的安全训练与预测。典型协议有结合秘密共享和混淆电路的 ABY3、XONN 等框架，结合同态加密、混淆电路、秘密共享等技术的 GAZELLE、Delphi 等框架。

开源框架

- TF-Encrypted 是由 Dropouts、Openmined、阿里巴巴参与组织的基于 Tensorflow 的隐私保护机器学习的开源框架。目前支持基于秘密共享技术的

Pond、SecureNN、replicated secretsharing 等三方安全计算协议，支持逻辑回归、线性回归等常见算法。

- CrypTen 是由 Facebook 公司开源的一个基于 PyTorch 的隐私保护分布式深度学习开源框架。底层协议采用算数秘密共享与布尔秘密共享的混合共享机制，支持神经网络推理和训练。

- Rosetta 是由矩阵元开发的基于 Tensorflow 的隐私 AI 开源框架，基于椭圆曲线算法、全同态加密算法、秘密分享和不经意传输算法、门限密码学综合应用的安全多方计算技术，底层协议为 SecureNN 和自研的 Helix 协议，可支持联合查询、联合建模、模型训练等。

- SyMPC 是由 Openmined 开源的一个支持 PyTorch 的隐私保护机器学习开源框架，底层支持 ABY3、Falcon、FSS、SPDZ 等多方安全计算协议，支持卷积、线性等神经网络层。

- SPU（Secretflow ProcessingUnit）是蚂蚁集团开源的隐语平台的密态计算单元，为隐语提供安全的计算服务。支持大部分 Numpy API，支持自动求导，提供 LR 和 NN 相关的 demo，支持 pade 高精度定点数拟合算法，支持 ABY3、 Cheetah 协议。

除了上述的几款开源框架外，还有复旦大学开源的 FudanMPL、原语科技研发的 Primihub、微软开源的 EzPC 等开源框架也是基于安全多方计算技术的隐私保护机器学习。

厂商

国内诸多厂商利用多方安全计算来增强 AI 数据的安全性和隐私性的，包括蚂蚁集团、阿里巴巴、百度、华控清交、原语科技、矩阵元等科技公司。

- 阿里巴巴摩斯多方安全计算平台是大规模商用的隐私计算产品，解决企业数据协同计算过程中的数据安全和隐私保护问题。目前产品已广泛应用于联合营销、政务数据安全开放、联合风控、多方联合科研等业务场景。

- 华控清交研发的 PrivPy 多方安全计算平台，实现了支持通用计算类型、高性能、集群化和可扩展的解决方案。支持标准的 Python 语言和 SQL 操作，兼容 NumPy 和 Pytorch 等函数库，能够支持包括绝大多数机器学习算法在内的计算类型和系统实现，支持联合模型训练和 AI 安全预测功能。

- 富数科技研发的 FMPC 安全多方计算产品支持私有化部署，通过秘密共享，混淆电路，同态加密等多种技术，实现安全多方求交、安全统计、安全矩阵运算等多种算子，便捷高效安全实现跨机构联合统计决策，多元数据分析等应用，适用于金融、医疗、政务、工业等多种场景。

（3）同态加密

同态加密是一类特殊的加密技术，是指对加密数据进行处理得到一个输出，将此输出进行解密，并保证该解密结果与同一方法处理未加密原始数据得到的结果一致。同态加密可以保障 AI 数据在计算过程中的安全性。

目前主要有三种同态加密方式：部分同态加密、类同态加密和全同态加密。

1. 部分同态加密：是指支持单一乘法或加法的加密方案。如支持乘法同态 RSA 和 ElGamal 算法，支持加法同态的 Paillier 算法。

2. 类同态加密：是只支持有限次加法和乘法运算的加密方案。

3. 全同态加密：是支持任意算法，且执行运算次数不受限制。目前全同态加密的计算复杂度通常要远高于部分同态加密。

目前同态加密技术在 AI 中的应用多集中在模型的预测阶段，而较少应用于训练阶段。这主要是因为训练阶段需要大量的矩阵运算和模型参数迭代更新，对计算性能和吞吐量要求非常高。而现有的同态加密算法，不管是部分同态、类同

态、全同态加密算法，都会导致显著的计算开销。此外，同态加密数据表示范围和精度有限，多次计算会累积误差。这将严重拖慢训练过程,使之难以满足实际需求。

同态加密方案虽然安全可靠，但只支持加法和乘法等多项式运算，而不支持机器学习过程中使用的非线性运算，如神经网络中的 sigmoid 和 ReLU 等激活函数 。解决方法主要分为两类：

1. 无需多项式近似的同态加密隐私保护机器学习方案：依赖数据持有者与模型所有者交互完成非交互线性激活函数的计算问题。

2. 基于多项式近似的同态加密隐私保护机器学习方案：利用多项式逼近技术来对激活函数等进行模拟，这种近似会造成精度和效率上的下降。

同态加密开源库

- SEAL 是由微软于 2015 年开始开发和维护的同态加密库，使用 C++开发，支持 BGV、CKKS 等同态加密方案，可集成到隐私保护的机器学习应用中。

- HElib 是 IBM 开源的一款流行的全同态加密开源代码库。目前实现的方案是包括带有引导的 BGV 方案和 CKKS 的近似数方案的实现,该框架使用了多种优化技术使同态运算更快。

- OpenFHE 是由 DARPA 资助支持的一个社区驱动的全同态开源项目，支持大部分主流全同态密码方案,包括 BFV、BGV、CKKS、 DM (FHEW)和 CGGI (TFHE)等。

- Hehub 是原语科技推出的国内首个自主研发的同态加密开源库，是一个易于使用，可扩展性强且性能优秀的密码学算法库。目前包含了 BGV、CKKS、TFHE 等全同态加密算法。

开源框架

- TenSEAL 是 Openmined 开源的一个支持同态加密隐私保护机器学习的 python 开源库。该开源框架底层基于微软 SEAL 同态加密库实现，提供端到端的加密机器学习，提供类似 tensorflow 或 pytorch 的高级 API，可以简化同态模型的训练和预测，支持各类机器学习模型转换为同态版本，包括逻辑回归、神经网络等。

- TF-SEAL 是基于 TensorFlow 和微软 SEAL 同态加密库的一个开源机器学习隐私保护框架，支持 BFV、CKKS 等多种同态加密方案，实现多种机器学习模型的同态版本，如线性回归、神经网络等。

- Concrete-ML 是由 Zama 公司提出的一种建立在全同态加密库 Concrete 之上的开源隐私保护机器学习工具集,旨在简化同态加密在机器学习隐私保护方面的应用。目前只支持模型的推理应用，可将逻辑回归、神经网络等模型转换成同态版本。

厂商

现有同态加密隐私保护机器学习解决方案效率或准确度损失较大，大规模商业应用较少。

- 阿里提出了一个猎豹（Cheetah）的新型框架，将全同态加密应用于深度神经网络的两方推理计算。

（4）差分隐私

差分隐私是由 Dwork 等人提出的一种建立在严格数学理论基础之上的隐私定义。通过差分隐私技术数据分析者能够获取有用的统计信息，而无法获得个体用户的敏感信息。

差分隐私机制是目前 AI 数据保护的最常采用的技术之一，主要通过添加一定的随机噪声（通常为拉普拉斯噪声或高斯噪声）来保护机器学习过程中模型和数据的隐私安全，以防止攻击者恶意推理。与加密方式相比较，差分隐私仅通过

添加噪声来实现，更易在实际场景中部署和应用，且不存在额外的计算开销，但是在一定程度会影响模型的精度。

机器学习中差分隐私扰动方法可以分为以下四类：

1. 输入扰动：输入扰动是指在模型训练开始之前，对个人数据先进行一定程度的随机扰动，保护个人数据敏感信息的泄露。一般分为全局差分隐私和本地化差分隐私机制。全局差分隐私，是指个人数据被集中收集后，通过差分隐私对全局数据进行扰动，可以看作训练数据的预处理过程。本地化差分隐私，是指个人现在本地对数据进行扰动后，上传给数据收集者。

2. 中间参数扰动：是指在模型训练过程中对梯度参数或特征参数引入噪声来防止敌手获取模型或训练数据的隐私。

3. 输出扰动：是指模型训练结束时，对模型权重参数进行扰动或对模型预测结果添加噪声，可以防止敌手对模型进行成员推理攻击或模型逆向攻击。

4. 目标扰动：是指机器学习模型的目标函数或目标函数展开式的系数中添加噪声。这种方式可以提高效率和可用，但是只能针对特定的目标函数。

开源框架

- Tensorflow Privacy 是谷歌发布的一个专注于机器学习中实现差分隐私保护的开源模块，提供包含高斯、拉普拉斯等多种差分隐私机制的实现，支持常见模型的差分隐私版本，如线性回归、逻辑回归、神经网络等。

- Opacus 是 PyTorch 的一个开源差分隐私模块，支持 DP-SGD、DP-Adam 等差分隐私优化算法，支持常见模型的差分隐私训练，如 CNN、RNN、GAN 等。

- OpenMined 的 PyDP 库是一个实现差分隐私的开源库，提供包括高斯、拉普拉斯多种差分隐私机制的实现，支持计算隐私损失，可以应用于机器学习模型。

- 隐语 SecretFlow 中提供差分隐私保护，支持 RDP 和高斯两种差分隐私策略，目前主要应用于联邦学习中。

厂商

差分隐私技术已被谷歌、苹果、微软、字节跳动、阿里巴巴等一些 IT 公司应用。

- 谷歌键盘（Gboard）中利用差分隐私训练语言模型；
- Microsoft Windows 将差分隐私应用于对使用和错误统计数据的收集中；
- 苹果利用差分隐私计算识别最流行的表情符号、最佳 QuickType 建议和 Safari 中的能耗率等内容；
- 字节跳动依托自研的 Jeddak 数据安全隐私计算平台，利用差分隐私来保障数据统计查询和用户数据采集过程中用户隐私。

此外，国内的一些厂商也将差分隐私应用于其隐私计算平台之中，如洞见科技的 InsightOne 隐私计算平台，百度 PaddleFL 联邦学习平台、腾讯的 Angel PowerFL 联邦学习平台等。

（5）可信执行环境

可信执行环境（Trusted execution environment，TEE）是在计算平台上由软硬件方法构建的一个安全区域，保证在安全区域内部加载的代码和数据在机密性和完整性方面得到保护，各方数据统一汇聚到该区域内进行计算，通过其安全特性提高终端系统的安全性。TEE 概念源于 2006 年提出 Open Mobile Platform（OMTP），是一种保护移动设备上敏感信息安全的双系统解决方案。在传统系统运行（Rich Execution Environment，REE）之外，提供一个隔离的安全系统用于处理敏感数据。2010 年 7 月，Global Platform

（致力于安全芯片的跨行业国际标准组织，简称 GP）起草指定了一整套可信执行环境系统的体系标准，成为当前许多商业或开源产品定义其各种功能接口的规范参考。

开源框架

lOP-TEE 是 Linaro 提出的基于 Trusted OS 的开源实现。OP-TEE 包括安全世界操作系统（OPTEE_OS），普通世界客户端（OPTEE_Client），测试套件（OPTEE_Test/XTest）和 Linux 驱动程序，该项目已经适配支持 28 多个平台/处理器；

lOpen TEE 是芬兰赫尔辛基大学和英特尔安全计算合作研究机构合作的项目，包含了若干软件工程，其中一个工程是 OP-TEE OS，还有实现 TEE 标准和机制的所需要的其他工程；

lTeaclave/MesaTEE 是百度提出的开源项目，是 Apache 孵化项目之一。其设计思路是构建一个类 FaaS（Function as a Service）的计算平台服务。平台在提供 TEE 机密计算、远程验证、安全存储等功能基础上，再通过一套任务管理框架实现了多任务的管理和并发操作；

lOcclum 是蚂蚁金服提出的开源项目，对应用代码不做更改或者只做少量调整，就可以迁移到 SGX 中运行，获得机密性和完整性保护。与其他 LibOS 项目相比，具有 Enclave 内多进程管理、全类型的文件系统支持、内存安全、容器化设计功能。

厂商

TEE 技术已经取得了一些进展和成果，当下全球头部的计算芯片公司都已经大规模商用了 TEE 架构方案。

lARM、Intel 和 AMD 公司分别提出各自的可信执行环境技术 TrustZone、Intel SGX 和 AMD SEV 及其相关实现方案；

- 诺基亚和微软整合的 TEE 框架称为 ObC，目前已经部署在诺基亚流光设备上；
- 三星的 TEE 框架名为 TZ-RKP，已经部署在三星的 Galaxy 系列设备上；
- 国产化 GPU 厂家依托于信创浪潮也在 TEE 方向研发创新，鲲鹏、飞腾、海光、兆芯等都已经在量产的芯片里提供了自主可控的 TEE 实现。

（6）其他传统数据安全技术

通过数据分类分级、数据标记、数据脱敏等多种技术对数据进行梳理，识别敏感数据并进行标记、分类分级以及脱敏、去标识化处理，有效降低敏感数据泄露风险。

数据分类分级通过应用机器学习、模式聚类、自然语言处理、语义分析、图像识别等技术，提取数据文件核心信息，对数据按照内容进行梳理，生成标注样本，经过反复的样本训练与模型修正，可以实现对数据自动、精准的分级分类。

数据标记是指对需要保护的数据增加标记信息，是实现数据分类分级安全防护的基础。通常分为分离式标记和嵌入式标记两类，分离式标记是指标记信息和原始数据分开，只建立两者间的映射关系，主要通过扩展元数据信息或者数据库表结构、建立索引表等方式实现，适用于数据访问控制、加密等场景；嵌入式标记是指将标记信息和原始数据融合形成新的带有标记信息的数据，主要通过密码标识、数字指纹、数字隐写等技术实现，适用于数据审计和溯源等场景。

数据脱敏是在不泄露敏感信息的提前下保留数据源的可用性，结合数据合规性规则智能生成脱敏特征库，并与敏感数据识别智能关联，实现智能发现和自动

脱敏，有效降低敏感数据泄露风险。数据脱敏技术分为静态脱敏和动态脱敏两种应用模式，静态数据脱敏技术一般是通过脱敏算法，将生产数据导出至目标存储介质，可以支持源库脱敏、跨库脱敏、数据库异构脱敏、数据库到文件脱敏、文件到数据库脱敏、文件到文件脱敏等场景。动态数据脱敏通过解析 SQL 语义匹配脱敏条件，通过改写或拦截 SQL 语句，返回脱敏后的数据到应用端，支持实时运维管理、应用访问等场景。

数据脱敏开源工具

ldeidentify 是基于 Python 的数据脱敏工具，对结构化和非结构化数据进行脱敏，支持多种脱敏技术；

lgo-mask 包，通过结构体的 tag 功能可以脱敏结构体中的任何字段，实现对字符串类型、slice 类型、map 类型进行脱敏处理。

厂商

数据分类分级、数据标记、数据脱敏等数据安全技术已逐渐成熟，在国内外数据安全厂商如 IBM、Symantec、启明星辰、天融信、绿盟科技、深信服、世平信息等进行应用并商业落地，形成数据泄露防护、数据安全治理平台、数据分类分级、数据脱敏等多种数据安全产品。

2.2.3 AI 算法的安全

2.2.3.1 算法分类

常用的算法类型有专家系统、传统机器学习与深度学习。专家系统通俗来说就是由认可的专家共同制定计算规则；传统机器学习是运用可解释的数学公式进行推导预测；而目前以 AIGC 为代表的深度学习则是模拟人脑神经元进行学习与预测，通常不具有可解释性，但在解决问题时具有更好的实用性。

国家互联网信息办公室、工业和信息化部、公安部、国家市场监督管理总局联合发布，自 2022 年 3 月 1 日起施行《互联网信息服务算法推荐管理规定》，提到推荐算法技术，包括

个性化推送类：最为人熟悉的应用在广告、短视频、“大数据杀熟”场景；

生成合成类：例如文章生成算法、换脸算法；

排序精选类：例如搜索引擎广告排序算法、电商平台店铺排序算法；

检索过滤类：例如应用在各类公众搜索引擎、电商或内容平台的内部搜索引擎的过滤不良信息的搜索引擎；

调度决策类：例如网约车平台、外卖平台的订单匹配算法等。

2.2.3.2 算法黑盒问题

AI 算法“黑盒”问题是指由于在 AI 产品上的深度学习等主流算法模型内部结构复杂、运行过程自主性较强且人工无法干预等因素，在数据输入、模型训练、结果输出等方面出现运行机制难以解释，运行结果无法完全掌控等问题。由于算法的技术黑箱性，用户无法明确理解推理和计算过程，只能被动的接受模型给出的输出,导致 AI 在实际应用中将面临如下几大主要的安全威胁：

训练数据不足或缺乏导致安全缺陷难发现；

模型运行自主性及不可解释性导致运行过程难理解；

安全测试标准不统一导致产品安全难掌控；

技术手段探索不足导致安全监管难以到位。

算法的黑盒问题制约 AI 应用场景的广泛落地，特别是对于一些安全性要求较高的场景，例如在自动驾驶领域中，AI 模型给出错误决策导致汽车偏离正常轨道造成交通事故，若由于算法黑盒问题，应急响应人员将无法进行有效的溯源

分析、定位根因，因此也无法有效避免类似事件的再次发生。要加速 AI 在此类场景中的落地，必须要解决算法的可解释性问题。目前已有相关方法可用于 AI 算法模型的可解释性增强，按模型训练前后阶段可分为集成解释、后期解释。

1.集成解释

通过算法优化与模块化提高可解释性:在训练数据较少时，选择人类较易理解的算法来训练模型，如决策树算法模型； 对于复杂模型，可设计成相互独立的模块化结构，每个独立模块可单独解释从而提高整体模型的可解释性，深度学习里的 Attention 技术、模块化网络结构、概率模型（Probabilistic Model）均可以作为模型内部的的独立部件；通过特征工程提高可解释性：模型训练前针对训练数据进行分析处理，进行特征提取帮助模型更加有效的实现数据处理，同时也方便人们在后续的推理过程当中分析模型的决策依据。

2.后期解释

后期解释是在模型训练完成后，针对特定场景基于模型输出，分析哪些样本对模型输出起决定性作用或具有较高的权重。后期解释可分为直接可解释性分析和非直接可解释性分析。直接可解释性分析利用模型相关信息如中间特征、梯度或参数等，结合可视化技术验证样本权重。非直接可解释性分析通过对样本数据的某个特征进行选取或者修改，并观察分析修改后的模型输出，从而得到模型样本不同特征对模型输出结果的影响。

2.2.3.3 算法脆弱性问题

算法脆弱性主要体现在算法模型泛化能力较弱，在模型面临复杂的输入，或者面对真实世界的异常场景下，不能够做出正确的推理与决策。算法脆弱性风险主要表现在：

- 算法设计局限性，特征描述的局限、目标函数的偏差、计算成本的制约都是可能导致决策偏离预期甚至出现错误结果的原因

- 数据“投毒”易导致结果异常，算法本身无法识别出异常样本，攻击者通过修改、删除部分样本或加入精心设计恶意样本等操作，导致训练出的模型可用性和完整性遭到破坏；

- 模型保密性和稳健性易受威胁，攻击者利用样本迭代对目标模型进行查询，基于返回结果构建出相似模型，进而还原出模型内部信息，基于这个相似模型，构造出对抗样本，使之做出错误决策

- 数据不均衡易引发“偏见”，人工智能决策结果的准确性、客观性很大程度依赖于数据本身。数据本身分布偏差、技术人员本身对事物认识的“主观性”，导致人工智能决策结果往往出现偏差

算法脆弱性导致的模型鲁棒性的缺乏，在某些场景下，可能导致严重的安全事故，例如医疗机器人在面对异常情况下的错误决策可能严重威胁到病人的生命健康安全、智能驾驶如果不能正确应对异常情况可能导致严重的交通事故。算法脆弱性的增强技术主要有数据增强、对抗训练等。

- 数据增强

模型的判断与决策能力来自于海量数据的训练过程。数据是大模型的养料，对模型训练至关重要，客观公正、完整的数据集很大程度上影响了模型判断的准确性。数据增强技术是指加强训练数据的采集力度，使得训练数据尽可能覆盖各种异常情况，从而提高模型面对异常情况的判断能力。如在 AI 图像识别模型训练过程中，针对现有的图像数据集，通过调整亮度、分辨率、进行仿射变换等迅速生成大量新的样本并组合原样本形成新的训练数据集。基于新数据集训练出的图像识别模型将拥有更高的判断准确率。

- 对抗训练

基于对抗攻击算法生成大量对抗样本并作为训练数据集的补充，从而提高模型面对异常输入数据的判断决策能力。Goodfellow 等人提出的对抗生成网络作

为一种对抗生成模型，由生成器和辨别器组成。辨别器负责鉴别样本是真实的还是伪造的，生成器用于模拟训练样本数据的分布生成逼真的对抗样本，对抗样本与真实样本一道作为训练数据集。

2.2.3.4 算法监管与知识产权风险

- 算法监管风险

监管强度和复杂度上升，对监管、审核、评估、验证各算法提供者的算法机制机理、模型、数据和应用结果，鼓励优化规则的透明度和可解释性，但不强求企业全盘托出；建立对交通、教育、生产制造等重点领域的公共测试数据环境支撑行业监管，都对算法机制机理的审核对象、程序都提出了更为具体的要求。

- 模型知识产权风险

模型训练数据的采集以及算法模型的设计开发已经逐渐变成了企业的重资产投入，性能优良的算法模型、参数以及训练数据已经成为企业、研究机构的核心资产。由于存在巨大的经济效益，大模型势必成为攻击者的窃取目标，导致企业面临较大的知识产权风险：

1)模型文件窃取：攻击者利用系统漏洞、供应链攻击等方式直接窃取算法参数文件，针对直接窃取模型文件导致的风险一方面需加强基础设施安全防护，另外可重点对模型文件做进一步的加强防护，如模型文件加密、可信计算、机密计算等

2)模型窃取攻击:攻击者通过对黑盒模型进行大量查询操作，获取查询的输出并进行分析，以期达到窃取模型参数或者对应功能的目的。通过对特定用户的访问报文及请求行为的监控、分析，提取攻击特征，对异常的 API 访问请求进行限流或直接拒绝服务，从而达到防御模型窃取攻击的目的。另外也可以基于模型水印技术，在原模型的训练阶段嵌入特殊的识别神经元，该神经元针对特定的输入（水印秘钥）返回特定输出。在发现相似模型的情况下，可通过特定的输入输

出来判断该模型是否通过窃取原模型所得。目前，典型的深度神经网络模型水印技术主要有静态水印技术、动态水印技术和主动授权控制技术。

2.2.3.5 针对算法模型的攻击

- 后门攻击

与传统应用相似，AI 模型同样面临着后门攻击的威胁。当存在后门时，模型能正确响应正常输入，但在针对后门制造者的特定输入时，输出将变成攻击者预先设置的恶意目标，从而实现对模型的操纵。攻击者通过直接修改模型（如植入特定神经元）、或者进行数据投毒实现后门攻击。针对直接修改模型神经元的后门攻击，需要加强对训练环境的保护防止系统入侵导致攻击者直接篡改模型文件；针对数据投毒型的后门攻击防御方法主要包括数据预处理和模型剪枝，从而达到破坏模型中可能存在的触发器。

- 对抗攻击

对抗攻击是指攻击者在输入请求中加入一些人类无法感知的扰动，但是算法模型能够识别出这些扰动，并且这些扰动会影响模型的推理导致做出偏离预期的决策。对抗训练是针对对抗攻击的强有力的防御方法。在模型训练阶段，通过技术手段生产大量对抗样本，这些对抗样本和原样本一起作为深度神经网络的训练数据，训练出来的模型可以主动防御对抗攻击。

- 注入攻击

类似传统的 SQL 注入、命令注入，恶意用户或攻击者可以在请求参数中注入精心构造的请求来操作算法模型的输出。注入攻击可分为目标劫持攻击、提示注入攻击等。目标劫持攻击是在请求提示词中添加一些恶意指令，使模型忽略掉最初的任务而执行攻击者指定的任务。提示注入攻击经常用于污染模型输出生成不合规内容、在内容审查过滤中绕过审查机制等。

注入攻击可通过输入侧防御与输出侧防御来进行安全加固。在输入侧可以通过规则过滤或黑白名单机制，确保恶意提示关键词无法与模型进行交互，从而避免提示注入攻击；同时，对输出侧的内容进行实时检测，对违法违规内容进行过滤，确保数据输出的安全性。

2.2.4 AI 应用的安全

AIGC 技术（人工智能生成内容）的高速发展引起社会广泛性关注，在写论文、生成文案、作画、作曲等创作性工作上展现出替代人类劳动力的趋势。在图像领域，以 Stable Diffusion 为代表的开源模型和 Midjourney 为代表的商业创作平台甚至能生成专业摄影级图片；在文本领域，以 ChatGPT 为代表的智能应用也在不断地重塑人们的工作方式。虽然新技术的产生极大程度上降低优质内容创作门槛、提升了工作效率，但同时也引起人们对其潜在的内容安全、数据安全、隐私保护等风险的担忧。

一方面，提供 AIGC 服务的平台使用的基座模型用到了大量的数据进行训练学习，但是训练和生成的过程都是黑盒的，很难完全保证生成内容的可靠性和可解释性。在这种情况下，模型生成的内容除了会产生潜在的内容风险之外，还可能引发一些伦理，隐私等方面的法律问题。另一方面，AIGC 作为新兴技术在提升人类生产效率的同时，黑产也会利用这些技术升级他们的手法，使得像谣言、诈骗、虚假新闻等场景的内容生产成本更低，更难被普通人察觉。针对上述的 2 种 AIGC 的应用安全问题，下面分别从“AIGC 应用安全防御”和“深度合成内容检测”两个方面进行介绍。

AIGC 应用安全防御

随着社会各界对大模型应用的关注，大模型安全技术也成为了目前研究领域的热点。要确保大模型应用推向市场后的安全性，需要在大模型应用的生命周期的各个环节切入，提供系统化的技术方案来解决大模型的安全性问题。蚂蚁集团

的“天鉴”内容风控引擎在大模型安全防御的实践中，分别从模型训练，用户交互和内容生成等环节切入，搭建了一套体系化的技术框架，来保障大模型应用的可靠和安全。

（一）大模型安全护栏

在大模型的应用中，目前主要都是通过单轮或多轮交互的形式来驱动大模型进行内容生成和创作，在实际的场景下，我们发现恶意用户会通过构造风险意图来诱导模型生成涉及到伦理，隐私，意识形态，色情低俗等风险内容。对此，通过在交互层加入“护栏”模块，对用户意图的进行精细化理解，并针对不同的意图类型制定对应的干预策略，可以有效的提前避免风险内容的产生。 “护栏”的类型可以从风险域和应用场景 2 个角度进行划分，例如风险方面就可以分为内容风险、隐私安全、伦理安全等；应用场景就可以从应用类型（文生文，文生图）和应用领域（金融，医疗）进行划分。

（二）可控内容生成

大模型的幻觉和偏见问题是目前业界热议的话题，相关的解决方案随着各项研究的深入也在持续的迭代升级。大模型的“幻觉”主要是指模型会生成“看似流畅的表达，但并不符合事实”的内容，其本身不一定是有害的，但是在一些领域下，则可能造成严重的后果。例如在金融应用中，输出错误的咨询信息，可能会给用户带来资资损，在政策类的应用中，错误的政策信息反馈可能会导致民生问题。

大模型的幻觉和偏见形成的原因有很多，包括数据噪声的影响，训练数据的分布问题，生成算法的解码策略等等。在实践中，我们发现，当赋予大模型更多自由生成的空间时，幻觉和偏见的现象也会增多。为了提升生成内容的可控性，可以在大模型应用生成内容的过程中提供更多的事实性的参考信息，以提升最终生成内容的可控性。可以采用的方案就包括：

1、基于问答库的标准回复

通过对一些关键性问题提前生成交互文案，并进行审核入库，在用户问到相关的问题时，直接调用审核后的回复，就可以最大程度的保障生成内容的可控性。

2、基于文档库的检索增强回复

对于一些比较泛的领域以及时事的内容，很难提前完成问答库的构建，可以通过在模型生成回复的过程中加入对应的文档检索增强的模块，提前关联相关的事实信息后，再送入大模型进行内容生成，来对齐生成内容和事实信息的一致性。

3、生成指令引导

对于偏见性内容，可以通过在交互过程中加入辅助性的指令，来控制内容生成的倾向性，比如在文生图的应用中，通过在指令中加入和价值观相关的控制指令，就可以一定程度抑制生成内容中出现不良信息的概率。

（三）训练数据去毒

模型训练数据的质量和内在的不当内容，也是导致大模型在应用时生成风险内容的重要原因之一。因此，可以通过对模型训练样本的进行毒性标注，并设计针对性的训练任务，来辅助大模型对齐人类价值观，来系统性地提升生成内容的可靠性。在实践过程中我们也发现，当引入足够的去毒语料后，裸模型的风险率有很大程度的降低。

深度合成内容检测

AIGC 技术逐渐成熟使得信息伪造的成本也越来越低，例如通过图像合成和 OCR 对抗攻击技术进行证件伪造的违法行为层出不穷；基于 AI 换脸和语音合成技术的熟人骗局也已经屡见不鲜。在内容伪造的门槛越来越低的情况下，需要加强对伪造风险的防御能力的建设。相关监管机构也在制定包括 AIGC 图像识别，虚假信息检测等多项审核专项能力的标准制定，AIGC 内容检测未来也会作为内

容安全审核体系中的一个重要能力。

目前，AIGC 检测算法主要依赖数据驱动的分类方案，从技术方案上主要分为“大规模鉴伪数据的构建”和“深度合成内容检测算法”2 个部分。除此之外，为了抑制 AIGC 被滥用，各大平台也开始投入在“数字水印技术”的研究，以建设透明、可追溯的 AIGC 内容生态。

（一）大规模鉴伪数据构建

随着 AIGC 技术的开源化和平台化，伪造信息的内容越发逼真，多样性不断提升，这也驱动了学术界和产业界构建了多个数据集，来推动 AIGC 检测技术的研究和评估，例如，在图像领域的 Sentry-Image、GenImage，在音频领域的 SASV2022、ADD2022，在视频场景下的 DFDC、WildDeepfake。但是从实际应用的角度来看，当前积累的数据集还只是冰山一角，能覆盖的模型类型，生成方式，模态融合的形式等方面都非常有限。目前产研界也在探索通过构建统一的多模态内容生成框架，并结合实际应用场景，系统化、规模化地生成伪造内容数据，以解决数据集规模小，多样性匮乏，模态单一等问题。

（二）深度合成内容检测算法

单一模态的深度合成内容检测中，对可鉴别的数据特征的挖掘是鉴伪算法中的关键部分。例如，在图像鉴伪中，人物轮廓的边界，图像细节的清晰度，画面中物理逻辑的违背等特征的引入可以一定程度提升真伪检测的精度；在音频鉴伪中，可以引入包括时域信号、频域信号、CQT、能量、Pitch 等常用的声音特征，并将其输入至深度神经网络等模型中，以提取更高维度的特征信息来辨别深度合成的内容；同时当前也有很大一部分工作聚焦在提升检测模型的泛化性。例如，在表征方面学习一个与内容无关的能表示扩散模型，自回归模型，生成对抗模型等不同类型生成模型的通用表征，以提升检测模型的训练效率；在训练中加入对抗训练的方式来避免因退化、传播、攻击等引起的检测效果衰减等。除了单模态的检测算法，融合多模态特征来更好地来检测实际场景中的复杂数据，也是目前

在探索的重要研究方向。

（三）数字水印技术

数字水印可以认为是 AIGC 平台在防止自身提供的服务被恶意应用时的一种主动防御手段。通过提前对生成的内容嵌入数字水印，可以检测时精准地进行判别和身份追溯，定位恶意用户。现有的数据水印技术主要应用于版权保护和数字照片注释等场景。在 AIGC 应用的场景中，生成内容往往会经过多轮传播和二次编辑，需要对应的数字水印的算法有更强的鲁棒性和抗攻击性，这是相关的研究重点关注的领域。

2.2.5 AI 安全的运管

- 组织架构

多元共治。在全国范围内应成立人工智能行业监督管理机构，对人工智能技术行业进行业务指导和监督；利用专业层面的优势，将分类分级监管原则贯彻落实，根据人工智能的应用场景对算法进行分级治理，辅助监管部门完成算法备案、算法审计及算法问责等工作，在此基础上不断完善通用人工智能算法的开发、使用及治理。

企业自治。设立独立的人工智能安全管理和执行部门，主要包括确立职责明确的部门和负责人，明确各安全岗位的职责和考核机制，以及贯彻执行各项制度。

建立虚拟协同机构，在业务、研发、法务、财务等部门制定人工智能安全专员，负责企业人工智能安全管理策略制定以及相关流程在本部门的实施。设立人工智能安全监督机构，负责定期对人工智能安全制度执行情况、技术工具执行有效性等开展监督检查

- 管理制度

在国家法规政策和标准规范的指导下，出台企业组织 AI 安全总体策略，编制企业标准规范与技术指南；设计安全管理制度、管理办法、安全操作流程以及运行记录单等;应急响应方面,主要包括确定一系列对不同情况的应急响应预案、准备应急技术工具和方案，并开展定期的应急响应演练等。

从企业层面整体设计人工智能安全制度体系，包括人工智能安全总纲，人工智能安全管理制度和，办法。人工智能安全总纲明确企业人工智能安全管理的目标、策略、基本原则和总的管理要求。人工智能安全管理制度和办法，是指人工智能应用全生命周期阶段的安全制度要求。面向人工智能应用各生命周期的安全操作流程规范是对人工智能安全管理办法的享系解释和补充,便于执行者理解和执行。

- 人员运营

人员运营方面，主要包括确定人工智能安全从业人员的招聘与聘用、培训与能力提升；对业务人员在人员管理、业务功能、机制设置、伦理合规和成本效率等方面进行培训；从人工智能安全管理、人工智能安全运营、人工智能安全技术等方面提升人工智能从业人员的能力。人工智能安全管理能力要求人员能够依据我国法律法规和兴业监管政策要求以及企业所属行业特点,制定企业人工智能安全策略，编制安全智能流程规范，实施人工智能安全应急指挥和跨部门管理协调能力，人工智能安全运营能力是指，在企业内部持续落实人工智能安全制度，通过管理手段和技术工具，对人工智能风险进行监测、识别、预警和处置。人工智能安全技术能力是指，跟踪和掌握人工智能安全前沿技术及发展趋势，熟悉主流人工智能安全产品和工具，能准确判定企业所需的最佳技术工具和具体实践。

3 AI 安全的热门问题

3.1 大模型安全

ChatGPT 引爆的生成式人工智能热潮，让 AI 模型在过去几个月成为行业瞩

目的焦点，并且在国内引发“百模大战”，在大模型高速发展的同时，大模型应用所面临的安全挑战、与潜在的威胁也不能够忽视。

（一）大模型的安全风险

首先，大模型在许多应用场景中处理大量敏感数据和个人信息，如用户的搜索记录、社交媒体互动和金融交易等。这使得数据泄露和隐私侵犯的风险不容忽视。一旦这些敏感信息遭受泄露，个人隐私权益可能会受到严重损害，甚至被用于恶意行为，如身份盗窃、诈骗和社会工程攻击。这不仅会对受害者造成经济损失，还可能导致社会的恐慌和不信任。其次，大模型的强大能力也可能被用于进行各种形式的恶意攻击。模型的对抗性样本攻击，即针对模型的输入进行微小改动，从而欺骗模型产生错误预测，已成为一种常见的威胁。恶意使用者可以通过这种方式制造虚假信息，影响决策结果，如将误导性的信息传播到社交媒体平台，从而扰乱社会秩序。再次，大模型的生成能力也可能被用于生成虚假的内容，威胁到媒体的可信度和新闻的真实性。另外，模型本身也可能成为攻击者的目标。模型参数和权重的泄露可能导致知识产权的损失，甚至使恶意使用者能够复制或修改模型，进一步恶化风险。对模型的针对性攻击，如投毒攻击，可能使模型的输出产生不良影响，从而影响到正常的业务运行。这些威胁可能在不经意间对企业和社会造成巨大的损失。最后大模型的使用往往涉及到社会伦理和法律问题。例如，算法的歧视性问题，即模型在处理数据时产生的不公平或偏见，可能引发社会的不满和争议。以及大模型可能会被用于传播虚假信息、仇恨言论或不当内容，从而引发社会不安定和文化冲突。

（二）大模型安全的应对建议

数据安全与隐私保护方向，大模型的技术原理决定了大模型安全应首先从训练数据做起。大模型生成内容的信息源是海量的语料库，也即训练数据。对大模型的数据安全与隐私保护可以从模型训练数据的全部数据处理活动出发。在数据获取阶段需保证训练数据获取的安全合规，对于自行采集的数据应确保数据的来

源合法，涉及到个人信息的应征的个人信息主体同意，如涉及敏感个人信息或者生物识别信息应获得单独同意；使用开源数据集的需要选择安全合规的数据来源并遵守开源协议，从第三方获取数据的应当与数据提供方签订合同、协议等措施确保数据提供方提供安全合规的数据；在数据存储阶段可以通过数据加密进行安全存储，从而提高数据的安全和隐私保护程度；通过限制访问对话记录的对象和权限，例如使用属性基加密对数据访问者的权限控制，或者使用代理重加密对用户自身的对话记录密文安全转换为有权限机构可解密的密文。利用密码学算法层面的访问权限控制手段、云计算中物理层面对访问权限设置等技术手段，限制可以访问对话记录的对象，起到保护用户隐私的目的；在数据传输、加工、提供等阶段可以通过联邦计算、差分隐私、数据沙箱、可信执行环境、数据沙箱技术入手，通过联合建模、数据隔离、随机噪声、密态计算、可信计算环境的技术构建确保用户数据、训练数据的安全；总之，在数据处理的各个活动中均需要实现数据安全和隐私保护。并且在所有活动中通过分类分级、元数据管理、数据流转审批、行为审计等管理手段确保数据安全及隐私保护。

模型安全保护方向,大模型在训练、管理、部署等各个环节需要重点关注训练语料的安全，模型资产的安全。训练语料需要实现高效的数据管理，防范模型原始语料数据泄漏，提高语料数据加工效率。可以通过元数据管理、数据分类分级、流转审批、数据鉴权、数据脱敏、加密保护、行为审计、访问控制数据备份回复等管理和技术手段实现模型训练语料的安全保护。模型资产需要保护大模型文件在训练、推理、微调等环节的模型文件安全。可以通过训练环境隔离、模型流转安全机制、推理和微调过程安全管控、模型私有化部署安全、模型审计与跟踪、模型安全修复机更新等管理和技术手段在大模型的整个生命周期中，确保模型资产的安全，保护敏感信息，防范恶意攻击，维护业务的正常运行。

内容安全方向，由于生成内容的复杂多样，合规要求逐渐完善，语言文化的差异以及内容识别和检测的多种难度，导致模型的输出内容安全保护成为行业内的重点关注问题。可以通过以下工作开展模型内容安全的保护工作：预训练数据

筛选，对训练数据进行训练数据进行筛选和清洗，删除偏见、不准确、以及违法违规内容保留高质量训练语料，同时对训练语料中涉及的个人信息、敏感信息等进行去标识化处理；输入内容干预，通过对输入内容进行人工审核、过滤技术或其他方式干预，以确保其符合特定的标准、规范和价值观，通过合适的内容干预尺度以及预置策略，兼顾大模型创新能力和回复内容的安性；提示词修改，通过对输入内容的识别、归类，将模型输入内容划分为为不同的主题类别和语义类别，对不同类别的和主题的输入内容进行检测并对不恰当内容进行高质量的提示词修改，可以有效提升大模型输出内容的合理性和准确性；大模型安全微调，基于已经通过安全审核的、符合安全标准的指令数据对大模型进行微调，通过 RLHF 方式提升大模型对安全回复内容的偏好程度，引导鼓励大模型生成更加高质量的安全内容；输出内容审核与改写，对大模型生成的文本内容进行检测和筛选，以识别并过滤掉有害、不准确、不适当或不合规的回复内容，可以通过红线知识库、多模态生成内容审核等技术手段实现。

此外，大模型在交互场景中的业务运营中，面临着多重安全威胁和风险，在账号安全、接口防刷、人机识别、AIGC 盗爬识别、设备风控以及风险情报等方面也应采取多种技术手段保障业务运营安全。综合运用这些措施，可以减轻大模型交互场景中的各种安全风险，保护用户隐私和数据安全，维护业务的稳定运行，同时，持续的监控、分析和改进也是确保业务安全的重要环节，以适应不断变化的安全威胁。

3.2 对抗样本攻击

描述

攻击者对输入样本进行难以察觉的微小的、有针对性的修改，使其几乎无法察觉，但对 AI 大模型能够引发误判或误分类，欺骗大模型并导致错误的输出结果。

攻击流程如下：

1、攻击目标的选择：攻击者首先选择一个目标 AI 大模型，该模型可能是一个自然语言处理模型、图像分类器、语言识别器等大模型。

2、收集训练数据：攻击者使用公开或其他手段获取的数据集用于训练目标模型。

3、生成对抗样本：攻击者使用训练数据以及例如：梯度优化方法、进化算法、生成对抗网络等对抗样本生成技术生成对目标模型的对抗样本。

4、评估对抗样本：攻击者对生成的对抗样本进行评估，以确保数据在人眼看来与原始样本相似，但可以欺骗目标模型。

5、攻击测试：攻击者尝试不同的对抗样本输入目标模型，并观察模型的输出结果，找到最有效的攻击方式。

6、攻击模型：找到最有效的攻击方式后，攻击者将样本用于目标模型的实际应用场景，以欺骗目标模型并导致错误的输出结果。

攻击带来的风险

1、误导模型：攻击者通过精心设计的对抗样本欺骗模型，使其将图像、语音或文本等输入误分类或误判，可能导致模型在实际应用中产生严重错误，影响决策的准确性和可靠性。

2、信息安全隐患：攻击者可以利用对抗样本来绕过模型的安全检测机制，欺骗系统进行非法操作或进行未授权访问，可能导致数据泄漏、模型窃取或其他的安全问题。

3、可信度破坏：当用户或系统无法信任模型的输出结果时，将会影响模型

的应用以及接受程度，最终导致用户对 AI 技术的不信任，减少对模型的使用和采纳，破坏了 AI 大模型的可信度和可靠性。

4、社会影响：对抗样本攻击可能对社会产生广泛并且严重的影响，如：在金融领域，攻击者可以使用对抗样本攻击来欺骗风险评估模型，导致错误的信用评估和投资决策，从而对经济产生负面影响。

防御手段

1、强化模型的鲁棒性：通过增加训练数据的多样性、使用对抗训练等方法来提高模型的鲁棒性。

2、对抗样本检测技术：使用比较输入样本和原始样本之间的差异或者通过监测模型输出结果的一致性等方式开发对抗样本检测算法，识别输入数据中的对抗样本。

3、输入数据的预处理：使用去噪、平滑化等技术来减少对抗样本的干扰。

4、集成多个模型：使用集成学习的方法，结合多个不同的 AI 大模型进行决策，使得攻击者需要同时欺骗多个模型才能成功攻击系统，降低对抗样本攻击的成功率。

3.3 数据投毒攻击

数据投毒攻击简介

数据投毒是指有意地向训练数据中注入恶意或欺骗性的样本，以干扰模型的性能或导致模型产生误导性的输出。在计算机视觉领域中，可通过污染训练数据来欺骗计算机视觉模型，使其在现实世界的应用中产生错误的结果。在生成式人工智能领域，可以干扰模型的生成结果、使其性能下降、生成具有误导性的内容、获取敏感信息或产生有害输出。

数据投毒攻击常见攻击技术

数据投毒主要有以下的攻击技术：一、数据偏斜，训练数据集中不同类别的样本分布不均匀，某些类别的样本数量过多或过少。可能导致模型对于少数类别的学习不足或过度关注多数类别，从而影响模型的性能和公平性。二、标签错误，将训练数据集中正确的标签标注为错误的类别。可能导致模型在训练过程中学习到错误的标签-特征关联，从而产生误导性的输出。三、标签矛盾，在训练数据集中故意引入标签矛盾的样本，如将相同的样本标注为不同的类别。可能导致模型混乱，难以准确地学习和分类。四、模型退化攻击，训练数据中包含大量垃圾数据，如利用模型生成的而非现实存在的数据，可能会造成模型的准确性下降、泛化能力降低、训练收敛困难或模型输出的不一致性增加的情况。五、后门攻击，模型训练时，将特定的标记、词语、句子或语法结构作为触发器。正常输入并不影响模型的性能，但当输入中包含了触发器样本时，模型会执行与预定行为相关的操作。这些操作可能包括将输出结果指向攻击者指定的结果、执行恶意代码或泄露敏感信息等。六、权重文件污染。通过在模型的特定层中注入恶意代码并生成权重文件，使用户在加载并进行推理时，自动触发注入内容。这种行为可能会导致攻击者能够在执行模型推理过程中执行任意代码，从而对系统造成危害。七、对抗样本攻击，通过对输入文本进行微小的、有针对性的修改，使模型在预测或分类任务中产生错误的结果，或者导致模型性能大幅下降。如使用贪心搜索算法生成的通用攻击后缀，能够突破生成式人工智能的道德、法律限制。八、反馈武器化，通过系统的反馈机制或算法的特性来实施攻击或滥用系统。如恶意差评、虚假反馈，从而误导、操纵甚至破坏或剥夺系统的正常功能。

数据投毒攻击的影响和风险

数据投毒主要有以下的风险。一、误导性结果，数据投毒可能导致系统做出错误的判断，或生成虚假、带有偏见或歧视性的内容，对信息的可信度和社会信任度产生负面影响； 二、信息泄露，数据投毒可能导致模型敏感信息被泄露。例如攻击者可能获取数据集中包含的隐私信息、商业机密，从而威胁隐私和数据

安全； 三、安全风险，数据投毒可能对个人隐私和公共安全构成威胁。例如攻击者可以通过欺骗人脸识别系统来绕过身份验证，或使用大模型进行网络攻击，甚至通过误导自动驾驶系统来引发交通事故； 四、法律和合规风险，数据投毒可能导致 AI 系统违反法律法规或合规要求。例如在金融领域，攻击者可能通过操纵数据来进行欺诈活动，从而触犯金融监管机构的规定。

数据投毒攻击的检测跟防御

数据投毒攻击主要有以下集中检测和防御方式。一、检查数据集中的标签、特征和其他元数据的一致性，如是否存在标签错误、数据偏斜或标签矛盾等问题。二、通过统计学、聚类分析和基于深度学习的方法，对数据集进行异常检测，识别数据集中不同机制产生的观测数据。三、对数据源进行可信度和完整性验证。规范数据采集过程，确保数据来源可靠，排除数据被篡改或污染的可能性。四、建立完善的模型监测机制，监测模型的输出结果并与预期输出做比较，当存在明显差异时应发出警告。五、使用对抗性训练技术来训练模型，增加模型的鲁棒性来应对数据投毒攻击。六、采用多模型投票的方式来减少单个模型受到数据投毒攻击的影响，提高系统的鲁棒性和可靠性。

3.4 供应链攻击

供应链攻击简介

供应链攻击（Supply Chain Attack）是一种高级威胁，通常涉及攻击者试图通过渗透和利用目标组织或产品的供应链环节来渗透或破坏目标系统或数据的安全性。这种类型的攻击通常不直接针对目标，而是利用供应链的脆弱性或信任关系来渗透目标系统。供应链攻击可以对各种组织和行业产生重大影响，包括政府机构、企业和个人用户。

供应链攻击具有多方面的关键要点。首先，攻击目标的多样性是显著特征，

攻击者可以针对硬件供应商、软件开发商、第三方服务提供商和云服务提供商等不同目标，以获取对更广泛系统或数据的访问权。其次，攻击手法的多样性包括恶意软件注入、数据篡改、硬件植入后门以及对第三方库或组件的恶意修改等多种方式，攻击者精心设计这些方法以确保难以检测或防止。攻击者的动机也多种多样，可能包括间谍活动、盈利、破坏竞争对手以及政治动机等，但无论动机为何，供应链攻击通常都会对受害者造成重大损害。供应链攻击影响整个供应链的各个环节，涵盖设计、制造、分销、交付和维护等环节。信任关系是供应链攻击利用的关键因素，组织往往信任其供应商或合作伙伴，使得攻击者更容易渗透。此类攻击通常难以检测，因为可以在产品或服务的生命周期的任何阶段进行，并且攻击者通常会采取措施来隐藏其活动，使其不容易被察觉。为预防和应对供应链攻击，组织需实施综合的安全策略，包括对供应商的尽职调查、安全审计、持续监控和应急响应计划，同时在整个供应链中建立强大的安全控制措施。

供应链攻击常见攻击技术

供应链攻击构成了极具广泛威胁的安全风险，攻击者利用多样化的手法试图渗透供应链的不同环节，以获取未经授权的访问权限、篡改数据或破坏系统。这些攻击方式包括恶意代码注入、脆弱性利用、可执行文件篡改、恶意数据注入、数据篡改、软件后门和木马程序等。攻击者往往针对供应链中的脆弱环节进行攻击，例如不安全的下载源或软件更新过程，通过向合法软件包或更新中注入恶意代码或篡改可执行文件或库来实施攻击。这种恶意代码注入可能会在用户或系统执行文件时触发，导致严重的安全问题。此外，攻击者可能尝试在数据源或数据传输过程中注入恶意数据，干扰正常的数据流或篡改数据，从而混淆或操纵决策。还有可能通过在供应链组件中插入后门或木马程序来建立持久性的未经授权访问权限。攻击者可能会专门针对特定的供应链组件进行感染、篡改或替换，导致整个供应链的崩溃或不安全性。甚至，攻击者可能瞄准硬件供应链，植入恶意硬件以获取对系统的控制权。

举例来说，攻击者可能会利用脆弱的 AI 开发库，如 TensorFlow 或 PyTorch，

在模型训练过程中注入恶意代码。此类恶意代码注入可能会在 AI 模型运行时触发，导致安全问题。另外，攻击者也可能针对 AI 开发中常用的数据集，通过数据篡改来误导模型学习，影响模型的预测结果。这些攻击形式对于 AI 系统的安全和性能构成严重威胁。这些攻击形式严重威胁着信息安全和系统稳定性。因此，重要的是加强对 AI 开发过程中涉及的具体软件、组件和服务的安全防护，以确保整个供应链的安全和完整性。重点关注和加强对供应链安全来源的保护和防范显得尤为重要。

供应链攻击的影响和风险

供应链攻击对 AI 系统和组织产生的影响多方面而广泛。首先，数据完整性和机密性受到严重威胁，攻击可能导致数据篡改或泄漏，损害数据的完整性并暴露机密信息，进而可能导致错误决策和法规合规问题。恶意数据注入和数据污染可能导致模型偏差，降低模型预测的准确性和可靠性，对 AI 系统的性能造成负面影响。模型质量受损进一步加剧了这种影响，影响业务流程并降低系统可用性。此外，供应链攻击可能导致 AI 系统的不稳定性，增加系统崩溃或服务中断的风险，影响业务连续性和客户满意度。这一系列影响还可能引发连锁反应，包括生产力下降、客户失信以及法律和财务方面的严重影响，对组织造成长期的负面影响。

供应链攻击的检测和防御

供应链攻击的检测是确保组织能够及时发现潜在威胁并采取行动的至关重要的一环。首先，建立持续监控系统是必要的，通过实时监测网络和系统活动，包括网络流量、日志、事件、系统性能和用户行为，以捕获异常情况。其次，利用异常行为分析工具如安全信息和事件管理系统（SIEM），对系统和用户行为进行深入分析，以发现不寻常的模式或活动。网络流量分析工具也应用于检测异常的数据传输或通信模式，这可能暗示供应链攻击的发生。定期审查系统、应用程序和设备的日志文件，以寻找与供应链攻击相关的异常事件或访问，是日志审计

的一项重要任务。此外，订阅威胁情报源的信息并进行分析，有助于及时获取有关已知供应链攻击的警报和指导。员工的培训也不可忽视，他们需要了解识别社会工程学攻击的迹象，并明白不点击恶意链接或下载附件的重要性。行为分析、机器学习、数据分析和挖掘技术都可用于建立模型以检测异常行为和特征。审查与供应链合作伙伴和供应商的关系，确保其符合安全最佳实践，也是预防供应链攻击的重要举措。最后，建立实时警报系统和供应链事件响应计划，能够在发现异常活动时迅速采取必要的应对措施。

供应链攻击的防御至关重要，以确保组织和系统免受威胁。为此，一些常见的供应链攻击防御策略被广泛采用。首先，通过定期进行供应链安全审查，评估与供应链相关的所有组件、合作伙伴和供应商，确保他们符合安全最佳实践。其次，进行供应链风险评估，识别潜在威胁和弱点，以了解哪些环节可能受到攻击，从而采取加强防御措施。另外，仅使用可信赖的供应链组件、库和软件，并确保供应链合作伙伴采取适当的安全措施。在软件和应用程序开发过程中实施安全最佳实践，加强代码审查、漏洞扫描和恶意代码检测。对传输和存储的敏感数据进行加密以防泄漏。实施多因素身份验证和强大的访问控制策略，限制对关键系统和数据的访问。建立供应链事件响应计划，包括隔离受感染组件、恢复操作和通知相关方的步骤。通过员工培训强调社会工程学攻击识别和安全意识。实时监控系统和网络活动，并建立响应团队以迅速应对威胁。定期备份关键数据和系统，并测试恢复过程。遵守适用的数据隐私和法规，确保数据处理合法并保护用户隐私。最后，减少对单一供应链合作伙伴或组件的依赖，实现供应链多元化以减轻某一环节受到攻击的影响。这些综合防御策略有助于确保组织免受供应链攻击的威胁。

3.5 数据泄露攻击

数据泄露攻击简介

在我们所处的这个技术快速进步的时代，人工智能（AI）技术正以前所未有的速度发展和普及。然而，与此同时，它也带来了不小的数据泄露风险，不仅威胁到使用者，还可能对模型的拥有者产生深远的影响。首先，这种风险源于 AI 技术所依赖的算法模型和运行参数的潜在漏洞。其次，由于 AI 技术在处理和分析数据时常常涉及到商业秘密和个人隐私数据，这使得样本数据的保护显得尤为重要。

在实践中，完全避免数据泄露几乎是不可能的，特别是考虑到某些模型的输出结果向量可以被用作攻击的工具。在模型逆向攻击中，攻击者能够利用模型的输出结果和其他相关信息，不直接接触到隐私数据的情况下反推出用户的隐私信息。此外，还有成员推理攻击，其中攻击者可以通过分析模型的输出来精准判断某一具体数据是否被用于训练集，这进一步加剧了数据泄露的风险。

因此，深入研究人工智能数据泄露攻击及其防御策略已经成为一个刻不容缓的任务。这不仅是为了保护个人和企业免受经济损失和法律风险，更是为了确保 AI 技术能够在一个安全和可控的环境中持续健康发展。只有这样，我们才能够充分利用 AI 技术带来的好处，同时避免其潜在的危险和风险。

数据泄露发生时机

伴随可能存在的管理不善，AI 数据泄露可能发生在 AI 系统生命周期的任何阶段，包括：系统设计开发阶段、系统测试阶段、系统部署交付阶段、系统运营阶段、系统退运阶段等。从严格意义的攻防来看，AI 数据泄露主要研究的是系统运营阶段所面临的恶意攻击行为，即攻击者通过反复调用、查询模型，根据模型返回的信息还原算法模型、训练数据和运行数据，例如：算法模型，通过对模型的访问而获得的输出数据构建替代模型，从而获得与被攻击模型相同或相近的特定功能；隐私数据，通过成员推理、数据逆向还原、属性推理等方式获得模型使用训练数据和运行数据的某些信息，或推断输出结果。

数据泄露攻击技术

人工智能技术数据泄露风险常常被忽视，然而考虑到隐私数据在线上黑市产业中的热门程度，就不难理解为什么必须研究人工智能数据泄露攻击技术。常见的数据泄露攻击包括：成员推理攻击、模型逆向攻击和无倾向神经网络数据提取等技术。

成员推理攻击是一种先进且具有潜在危险性的网络安全威胁。在这种攻击策略中，攻击者首先努力收集或制定大量可能的查询数据，这些数据可以来源于公开的数据集或其他渠道。随后，他们依照目标模型的行为特点创建数个“影子模型”来模拟其运作方式。接着，攻击者利用类似生成对抗网络（GAN）的技术来培训一个鉴别器。这个鉴别器的任务是学习如何根据影子模型的输出来区分哪些数据是训练数据，而哪些是非训练数据。换句话说，它能够识别出某个数据样本是否曾被用于训练目标模型。这个过程完成后，鉴别器可以被用来针对目标模型进行攻击，以估算特定的数据样本是否已被用于训练该模型。令人担忧的是，这种攻击可以在“黑盒”环境中进行，即攻击者不需要完全了解或访问模型的内部结构和细节，他们只需依赖模型的输出来进行分析和推理。通过这种方法，攻击者可能能够获取到一些敏感信息，从而可能追溯到个体的私人数据，这将对个人隐私造成严重威胁。因此，这种攻击方法强调了在设计和部署机器学习模型时，必须采取适当的安全和隐私保护措施。下图给出了成员推理攻击过程的示意图 5：

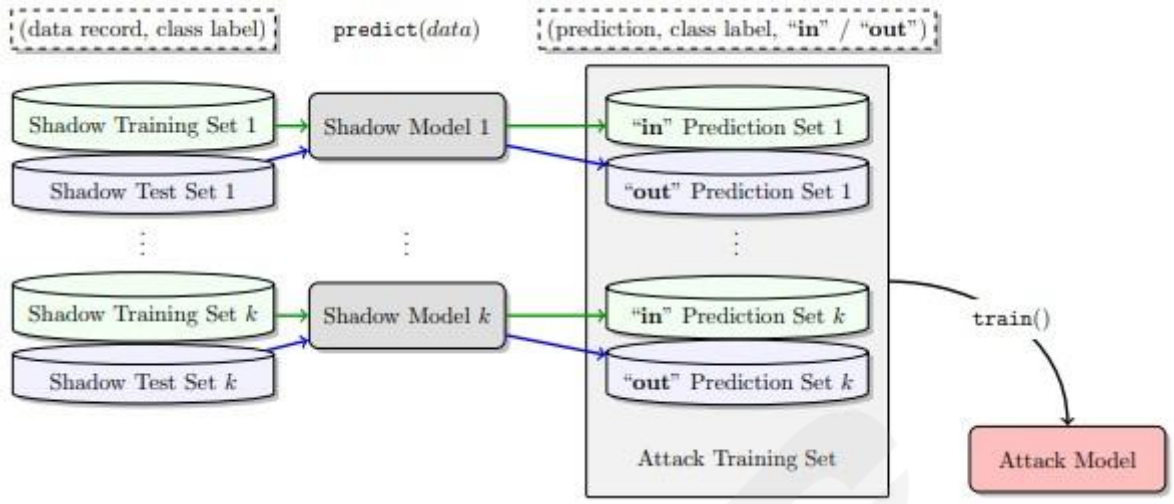


图 5 成员推理攻击示意图

数据泄露防御方法

AI 数据泄露攻击带来的经济风险和法律风险，使得人工智能技术的拥有者不得不认真思考针对数据泄露攻击的防御技术。近年来，涌现出了多种针对 AI 技术的攻击的防御机制。主要包括：更改模型结构、信息混淆和查询控制等方法。

更改模型结构：对模型结构做适当的修改，以此来减少模型被泄露的信息，或者降低模型的过拟合程度，从而完成对模型泄露和数据泄露的保护。例如：通过修改模型的损失函数，使目标模型给出的输出数据中包含尽可能少的信息。

信息混淆：对模型的输入数据或预测结果做模糊操作，在保证 AI 模型输出结果正确的前提下，尽可能地干扰输出结果中包含的有效信息，从而降低数据泄露的风险。信息混淆技术主要包含两类：1）截断混淆，对模型返回的结果向量做取整操作，抹除小数点某位之后的信息；2）噪声混淆，对输入样本或输出的概率向量中添加微小的噪声，从而干扰准确的信息。

查询控制：依赖使用者的查询记录，通过对样本特征或行为特征的分析，完成对数据泄露攻击的识别和防御。查询控制防御主要包含两类：异常样本检测和查询行为检测。异常样本检测利用攻击者通常会对正常的样本进行有目的地修改

这一特点，对正常样本的修改改变了样本的特征，样本特征之间的距离分布可以用来判断用户是否正在实施攻击；行为检测技术利用攻击者查询行为与正常行为会有较大不同，例如：访问的数量与频次等信息。模型提供者通过对用户查询次数、频率、时长等因素进行限制，从而提升攻击者的攻击成本。

3.6 模型窃取攻击

描述

模型窃取攻击是指攻击者通过访问模型的查询接口、观察模型的输出行为或使用生成的对抗样本等方法，来推断出模型的内部结构和参数，从而实现对模型的窃取。

攻击流程如下：

1、获得权限与数据：攻击者可能通过合法或非法手段获取到 AI 大模型的访问权限，包括模型的查询接口或训练数据。

2、模型输出探测：攻击者通过向模型输入一系列样本数据，并观察模型的输出结果。攻击者可能会探索不同的查询方式、输入数据的变化和模型的响应，以获取更多关于模型行为的信息。

3、推理模型结构：基于观察到的模型输出和相关的背景知识，攻击者试图通过机器学习、逆向工程或其他推理手段推断出模型的内部结构，包括模型的架构、层数、激活函数、参数等。

4、复制重建模型：使用模型结构与参数对模型进行重建或复制。

攻击带来的风险

1、竞争优势的丧失：通过模型窃取攻击使得攻击者能够获取目标模型的知识和技术，无需投入大量时间和资源来获得与目标模型相似的能力和性能，从而

可能导致模型拥有者的竞争优势丧失。

2、潜在的社会风险：攻击者可以使用窃取的模型进行恶意行为或非法用途，如：在金融领域，攻击者可以使用到风险评估模型来规避安全检测和欺骗系统，从而获得未授权的访问权限或进行欺诈行为，可能对个人、组织或整个社会造成严重损失。

防御手段

1、访问控制：使用身份验证、访问令牌、IP 过滤等机制，限制对 AI 大模型的访问权限，确保只有授权的用户或系统可以访问模型。

2、模型加密：可使用可信计算环境，如 TEE 技术，对模型以及输入&输出内容进行保护，防止以防止攻击者通过观察模型输出或分析模型参数来窃取模型。

3、对抗窃取检测：开发模型窃取检测算法，通过监测模型查询的模式、检测异常查询行为或比较模型输出的一致性来进行模型窃取的检测，以监测和识别潜在的模型窃取行为。

4、安全审计和监控：建立安全审计和监控机制，对模型的访问和使用进行监测和记录。及时检测和响应异常行为，以防止模型窃取攻击的发生。

3.7 AI 伦理/对齐

- 背景

过去十年中，人工智能取得了快速而显著的发展，逐渐渗透到各个学科和社会各个方面，如机器学习、自然语言处理和计算机视觉等技术的应用。人工智能已经广泛应用于商业、物流、制造、交通、医疗、教育、国家治理等领域，逐渐接管人类任务并取代人类决策。

然而，人工智能的发展也带来了一系列重大的伦理问题和风险。近年来，已经出现了许多人工智能产生不良结果的案例。例如，特斯拉汽车的一起事故中，自动驾驶系统未能识别迎面而来的卡车，造成司机死亡。微软的 AI 聊天机器人 Tay.ai 在加入 Twitter 不到一天后就变成了种族主义和性别歧视者，被迫下架。还有其他许多例子涉及人工智能系统的公平、偏见、隐私和其他伦理问题。更为严重的是，人工智能技术已经开始被犯罪分子用来伤害他人或社会，如利用基于 AI 的软件冒充某公司首席执行官的声音进行欺诈。

因此，解决人工智能带来的伦理问题和风险，使其在伦理规范的导向下发展和应用，变得非常紧迫和重要。人工智能伦理涉及的内容广泛，可以大致分为两个方面：人工智能伦理学和伦理人工智能。人工智能伦理学研究与人工智能相关的伦理理论、指导方针、政策、原则、规则和法规。伦理人工智能研究如何遵循伦理规范来设计和实现行为合乎伦理的人工智能。人工智能伦理学是构建伦理人工智能的先决条件，它涉及伦理或道德价值观和原则，决定了伦理道德上的对与错。只有有了适当的人工智能伦理价值观和原则，才能通过一些方法和技术来设计或实践伦理人工智能。

- AI 伦理问题的分类

根据 Changwu Huang 教授的观点，人工智能系统主要服务于个人或社会公众。因此，我们可以从个人和社会的角度分析和讨论人工智能伦理问题。同时，作为地球上的实体，人工智能产品不可避免地会对环境产生影响。因此，还需要考虑环境方面相关的伦理问题。因此我们可以将人工智能伦理问题分为三个不同层次，即个人、社会和环境层面的伦理问题。个人层面的伦理问题主要包括 AI 对个人及其权利和福祉产生不良后果或影响的问题。社会层面的人工智能伦理问题考虑了人工智能为群体或整个社会带来或可能带来的不良后果。环境层面的人工智能伦理问题关注人工智能对自然环境的影响。

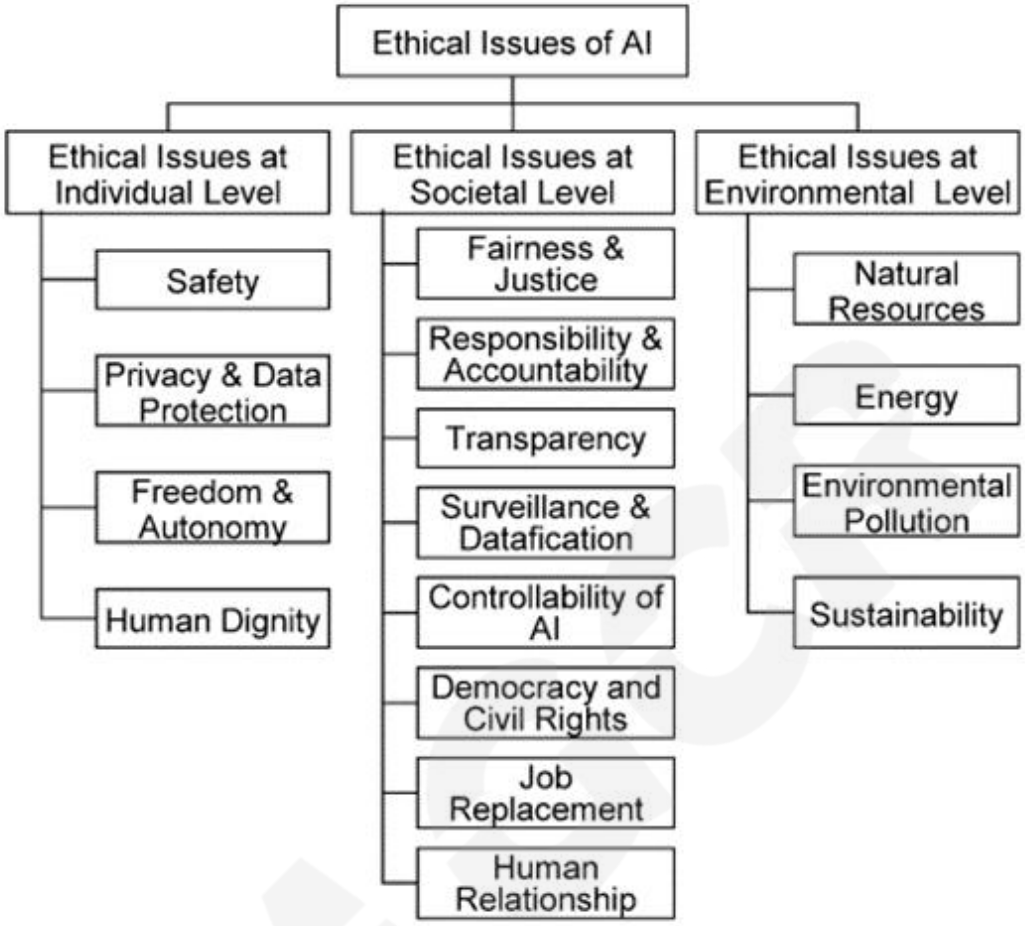


图 6 人工智能伦理问题

1） 个人层面的 AI 伦理问题

在个人层面，人工智能对个人的安全、隐私、自主权和人格尊严产生了影响。人工智能应用给个人安全带来了一些风险，例如自动驾驶汽车和机器人可能导致人身伤害事故。隐私问题也是人工智能带来的重大风险之一。为了提高性能，人工智能系统通常需要大量数据，其中可能包括用户的私人数据。然而，这种数据收集存在严重的隐私和数据保护风险。此外，人工智能的应用可能对人权产生挑战，如自主权和尊严。自主权是指独立、自由且不受他人影响的思考、决策和行动的能力。当基于人工智能的决策在我们的日常生活中被广泛采用时，我们的自主权可能受到限制。作为主要人权之一，人的尊严是指一个人受到尊重和以合乎

道德的方式对待的权利。在人工智能的背景下，保护尊严至关重要。人的尊严应该是保护人类免受伤害的基本原则，在开发人工智能技术时应该受到尊重。例如，致命的自主武器系统可能违反人的尊严原则。

2) 社会层面的 AI 伦理问题

在考虑社会层面的人工智能伦理问题时，我们主要关注人工智能为社会以及世界各地社区和国家的福祉带来的广泛后果和影响。在社会层面的伦理问题分类下，我们讨论公平与正义、责任与问责、透明度、监控与数据化、人工智能的可控性、民主与公民权利、工作替代与人际关系。

人工智能存在偏见和歧视，对公平正义提出了挑战。人工智能中嵌入的偏见和歧视可能会增加社会差距并对某些社会群体造成伤害。例如，在美国刑事司法系统中，用于评估犯罪风险的人工智能算法已被注意到表现出种族偏见。责任意味着对某事负责。为参与者分配责任对于塑造算法决策的治理非常重要。基于这一概念，问责制是对损害负有法律或政治责任的人必须提供某种形式的正当理由或补偿的原则，并通过提供法律补救措施的责任体现出来。因此，应建立机制以确保人工智能系统及其决策后果的责任和问责制。由于人工智能算法的黑盒性质，缺乏透明度已成为广泛讨论的问题之一。透明度，即对人工智能系统如何工作的理解，对于问责制也至关重要。监控和数据化是我们生活在数字化和智能化时代的所面临的问题之一。数据是通过智能设备从用户的日常生活中收集的，这导致我们生活在大规模监控中。随着人工智能的力量迅速增强，人工智能系统的发展必须要保障和确保人工智能系统的人类可控性。其他问题，包括民主和公民权利、工作替代和人际关系，也属于社会层面的问题。

3) 环境层面的 AI 伦理问题

环境层面的人工智能伦理问题关注人工智能对环境和地球的影响。人工智能可以为我们的生活带来很多便利，可以帮助我们应对一些挑战，但它也给地球带来了负担。人工智能的广泛应用往往需要部署大量的硬件终端设备，包括芯片、

传感器、存储设备等，这些硬件的生产消耗了大量的自然资源，尤其是一些稀有的元素。此外，在这些硬件的生命周期结束时，它们通常会被丢弃，这可能会造成严重的环境污染。另一个重要的方面是人工智能系统通常需要相当大的算力，这伴随着高能耗。此外，从长远和全球的角度来看，人工智能的发展应该是可持续的，即人工智能技术必须满足人类发展目标，同时保持自然系统提供经济和社会所依赖的自然资源和生态系统服务的能力。综上所述，人工智能发展所涉及的自然资源消耗、环境污染、能源消耗成本和可持续性是环境层面的主要问题和关注点。

1. 解决人工智能伦理问题的技术手段

通过新技术尤其是机器学习技术，以消除或减轻当前 AI 的缺点，规避相应的伦理风险。例如，对可解释机器学习的研究旨在开发新的方法来解释机器学习算法的原理和工作机制，以满足透明或可解释性原则。公平机器学习研究使机器学习能够做出公平决策或预测的技术，即减少机器学习的偏见或歧视。现有技术方法方面的工作主要集中在几个重大和关键的问题和原则上（即透明度、公平正义、非恶意、责任和问责、隐私），其他问题和原则很少涉及。因此，我们仅对涉及上述五项关键伦理原则的技术方法进行简要总结。

原则	包含领域
Transparency	Explainable AI or interpretable AI
Fairness & Justice	Fair AI
Non-maleficence	Safe AI Secure AI Robust AI

Responsibility & Accountability	Responsible AI
Privacy	Differential Privacy Federated or Distributed Learning Homomorphic Encryption

表 4 五原则

3.8 AI 辅助安全

一、简介

人工智能（Artificial Intelligence，简称 AI）正日益渗透到我们生活的方方面面，从驾驶到医疗再到客服，AI 辅助技术已经成为现代社会的重要组成部分。然而，随着 AI 技术的迅速发展，我们也面临着一系列与 AI 相关的安全问题。在 AI 辅助驾驶、AI 医疗辅助和 AI 智能客服等领域，确保 AI 系统的安全性至关重要。

自动驾驶汽车被认为是未来交通的一个重要发展方向，可以提高交通效率、减少事故，并改善行车体验。然而，确保自动驾驶汽车的安全性是一个巨大的挑战。AI 系统需要能够准确地感知周围的环境、做出适当的决策并控制汽车的行驶。任何一个系统错误或漏洞都可能导致严重的后果，甚至危及人们的生命安全。因此，确保 AI 辅助驾驶系统的安全性是至关重要的，包括对 AI 算法进行充分测试和验证，强调数据的质量和可靠性，以及加强网络安全防护。

AI 医疗辅助也是一个重要的应用领域。借助 AI 技术，医疗领域可以提高诊断准确性、优化治疗方案，并协助医生做出更好的决策。然而，在使用 AI 进行医疗辅助时，安全问题也需要引起我们的关注。不正确的诊断或治疗建议可能对

患者的健康产生严重影响。因此，确保 AI 医疗辅助系统的安全性至关重要。这包括确保 AI 算法的准确性和可信度，保护患者隐私和数据安全，以及监督和审查 AI 系统的使用。

AI 智能客服已经广泛应用于各个行业，为客户提供快速和个性化的服务。然而，虽然 AI 智能客服可以提高效率和用户体验，但也面临一些安全风险。例如，未经授权的访问和滥用个人信息的风险。确保 AI 智能客服系统的安全性涉及到保护用户隐私和数据安全，制定合适的数据使用政策，以及监测和防止恶意攻击。

AI 辅助安排不仅涉及到技术层面的安全措施，还需要政府、学术界和产业界的紧密合作，共同制定相关法规和标准，推动 AI 技术的可持续和负责任的发展。只有通过共同努力，我们才能充分利用人工智能的潜力，并创造一个安全可靠的 AI 时代。

二、Ai 辅助驾驶安全问题

AI 辅助驾驶技术的迅速进步正在汽车领域掀起一场革命，为我们提供了更多自动化和便捷，但与此同时，也带来了一系列潜在的安全问题。本文将详细探讨这些问题，借助实际案例来阐释它们的重要性和实际影响。

- 误识别和感知问题：AI 辅助驾驶系统的核心任务是感知和识别周围环境。然而，在某些情况下，这些系统可能出现误识别问题，未能准确辨认道路上的障碍物、其他车辆或行人，从而可能导致危险局面。在 2018 年，一起特斯拉 Model X 的致命事故发生在加利福尼亚。该车正在使用自动驾驶功能，但未能识别并规避一段道路上的损坏护栏，结果车辆撞上护栏，发生火灾，并造成驾驶员丧生。事故调查显示，车辆的感知系统未能正确辨识护栏，这导致了事故的发生。

- 数据质量和完整性问题：AI 辅助驾驶系统依赖高质量的地图和传感器数据，以进行实时的导航和感知。然而，如果这些数据不准确、陈旧或不完整，

系统可能会做出错误的决策。在 2019 年，一辆特斯拉 Model 3 在使用自动驾驶功能时碰撞到了一辆停在道路上的灭火器。事故调查揭示了一个问题，即灭火器并未包括在车辆的地图数据库中，因此系统未能识别它。这一案例突显了 AI 辅助驾驶系统对于准确和实时地图和传感器数据的重要依赖。

● 人机界面问题：人机界面是驾驶员与 AI 辅助驾驶系统之间的纽带，应该提供明确的信息和指导，以确保驾驶员了解系统的功能和限制。如果界面设计不清晰或混乱，可能导致驾驶员的误解或混淆，降低系统的安全性。有些车型的人机界面设计可能不够直观，导致驾驶员对系统的功能和限制了解不足。这可能导致驾驶员在需要时无法有效地介入，或者错误地过度依赖系统，增加事故的风险。

● 安全漏洞和黑客攻击：随着汽车的互联化，AI 辅助驾驶系统面临着黑客攻击的威胁。如果黑客能够入侵车辆的控制系统，他们可能能够远程操控车辆，威胁驾驶员和乘客的生命安全。在 2020 年，一名安全研究人员成功黑客入侵了一辆特斯拉 Model 3，通过漏洞远程控制了车辆的功能。这个案例引发了对汽车网络安全的关切，凸显了 AI 辅助驾驶系统可能受到网络攻击的风险。

● 过度信任问题：一些驾驶员可能过度依赖 AI 辅助驾驶系统，错误地认为这些系统可以应对所有情况。这种过度信任可能导致驾驶员在需要时无法迅速介入，尤其是在紧急情况下。过度信任问题在许多事故中都起到了作用，因为驾驶员可能过度依赖系统，忽视了对道路和其他车辆的监控。这种行为可能增加事故的风险。

● 法律和责任问题：涉及到事故责任时，AI 辅助驾驶系统引发了复杂的法律争议。确定责任可能会变得复杂，因为问题涉及到制造商、驾驶员和系统之间的责任分配。当发生事故时，制造商和驾驶员之间的责任分配可能会引发争议。如果系统出现故障或误识别，责任的划分可能会成为法律问题。

三、Ai 辅助医疗诊断安全问题

a.引言

在当今社会中，人工智能（AI）已经开始在医疗领域发挥越来越重要的作用。AI 医疗辅助诊断系统能够帮助医生更快速、更准确地诊断疾病，并提供个性化的治疗建议。然而，与 AI 的广泛应用相关的安全问题也不容忽视。本章将重点关注因 AI 问题导致的安全问题，探讨 AI 医疗辅助诊断的安全挑战和解决方案。

1. 数据隐私和保护

数据隐私和保护是指在处理和存储敏感数据时采取措施，以确保这些数据不被未经授权的访问、泄露或滥用。在 AI 医疗辅助诊断领域，数据隐私和保护至关重要，因为医疗数据包含患者的个人健康信息，如病历、诊断、药物处方等，这些信息必须严格保护。

- 隐私泄露：AI 系统需要访问患者的敏感健康数据。如何确保这些数据的安全和隐私成为一个重要问题。

〇案例 1：Bleeping Computer 网站披露，加利福尼亚州 Heritage Provider Network（全美最大的综合医疗服务网络之一） 中多个医疗机构遭遇勒索软件攻击，大量患者敏感信息泄露。[1]

〇案例 2：安全研究人员耶利米 •福勒（Jeremiah Fowler）在 Secure Thoughts 上发表文章称，他发现近 260 万条包含姓名、医疗诊断记录、保险记录和支付记录在内的个人病历数据被泄露了。福勒指出，不少被泄露的数据都指向一家名叫 Cense 的美国 AI 公司。[2]

- 数据滥用：存在滥用患者数据的风险，例如未经授权的数据访问或用于营销目的。

〇案例：雷峰网消息，谷歌在未经授权的情况下，非法使用了 160 万份 NHS

（National Health Service，即英国国家医疗服务体系）患者的机密医疗记录，并且将因此面临集体诉讼。[3]

- 数据匿名化：匿名化数据以保护患者隐私，但如何确保匿名化的有效性以及避免重新识别是挑战之一。

○案例：有一些成熟的平台和工具，可以实现数据的去标识化，从而保护患者的隐私安全。去识别化功能包括删除敏感信息，例如姓名和社会保险号，这些信息可能直接或间接地将个人与其个人数据联系起来。去标识化平台可以以极高的准确性匿名化文本内容中的敏感数据。[4]

2. 模型安全性

模型安全性是指确保机器学习和人工智能模型在其设计、训练、部署和运行过程中能够抵御各种安全威胁和攻击的能力。模型安全性是人工智能安全的重要组成部分，关注如何保护模型的完整性、保密性、可用性以及防止模型被恶意利用或攻击。

- 对抗性攻击：AI 模型容易受到对抗性攻击，攻击者可能通过修改输入数据来误导模型，从而导致错误的诊断。

○案例：据《纽约时报》3 月 21 日报道，科学家们开始担心，越来越多被应用到医疗保健服务中的 AI 技术，可能是一把双刃剑。除了帮助医生高效工作，它更有可能被蓄意操纵，导致误诊，以及其他更严重的后果。[5]

- 模型解释性：AI 模型通常是黑盒模型，难以解释其决策过程。这可能引发不信任和安全担忧。

○案例：当医生拿出一张 CT 扫描照片，向患者告知“根据人工智能算法的判断，您可能患病了”，病人会相信这一结果吗？是否需要医生进一步向患者解释，这一算法依赖于哪些参数，使用了哪些函数，是如何得出这一诊断结果的？

令普通人困惑的原因在于，输入的数据和答案之间的不可观察空间。这样的空间通常被称为“黑箱”（black box）。[6]

3. 数据质量和可信度

数据质量和可信度是关于数据的两个重要方面。数据质量是指数据的准确性、完整性、一致性、可用性和可靠性程度；数据可信度是指数据的可信程度和可靠性，它关注数据的来源、收集方法以及数据的背景信息。高质量和可信的数据对于建立可靠的模型和进行准确的分析至关重要。

- 噪音数据：医疗数据中可能存在噪音，包括错误或不准确的标签。如何处理这些数据以确保诊断的准确性？

○案例：近日，一篇发表在《自然》子刊《自然机器智能》的论文指出，华盛顿大学的研究人员对人工智能（AI）检测新冠病毒模型研究发现，这些模型存在不稳定性，可能会导致诊断失误的现象。关于误诊的原因，研究人员认为，主要是大部分模型只是依靠数据的分析和对于患者的胸片标注特征等数据，对患者是否感染新冠病毒进行判断，而不是根据真正的医学病理去诊断、分析。[7]

- 模型偏差：如果 AI 系统训练数据中存在偏差，可能导致不平等的诊断结果，从而引发道德和法律问题。

○常见的问题来自不均衡的数据集，比如在一个有关医疗的训练数据集中，某些人群没有样本表示。例如采用白种人患者数据. 进行训练的模型，可能在其他种族患者中效果不佳。[8]

4. 法规和合规性

法规和合规性（Regulations and Compliance）是指在任何领域，特别是在科技、医疗、金融和数据处理等领域，遵循和遵守相关法律、法规和行业标准的重要性。

- 医疗法规：AI 医疗辅助诊断需要遵守严格的医疗法规。如何确保 AI 系统的合规性是一个挑战。

○案例：医务人员在实际使用医疗 AI 产品过程中造成患者损害的，一般由医疗机构承担医疗损害责任，但如果上述医疗损害系由医疗 AI 产品本身的缺陷导致的，医疗机构有权向医疗 AI 产品的生产者或销售者进行追偿。[9]

○法规：《医疗事故处理条例》规定，由于医疗机构及其医务人员使用产品的过程中，违反医疗卫生管理法律、行政法规、部门规章和诊疗护理规范、常规，过失给患者造成人身损害的事故，医院应当向患者承担民事赔偿责任。情节严重的情况下，负有责任的医院和医务人员可能承担行政责任（如限期停业整顿、吊销执业许可等）甚至可能承担刑事责任（医疗事故罪）。[10]

- 透明度和报告：需要透明的报告机制来记录 AI 系统的决策和结果，以满足法规要求。

○案例：医疗机构使用 AI 系统进行诊断，但缺乏透明度和报告机制，医生无法了解 AI 系统的决策过程，也无法验证其诊断。[11]

5. 数据安全

数据安全是确保数据的机密性、完整性和可用性的过程和措施。数据安全旨在保护数据免受未经授权的访问、篡改、破坏或泄露的威胁。

- 数据存储和传输：如何安全地存储和传输医疗数据，以防止数据泄露或被恶意访问。

○案例：2020 年，医疗影像 AI 公司汇医慧影被黑客入侵，有消息提到，该公司的新冠病毒检测技术数据，正在被黑客以四个比特币（时价约合人民币 21.8 万元）的价格在线出售。之后，汇医慧影对此做出回应，称四月中旬，公司在境外公有云远程部署培训公益平台过程中，遭到了黑客的攻击。但黑客盗取的仅是

培训资料，没有 AI 源代码，更没有客户数据。[12]

- 数据备份：数据备份和灾难恢复计划对于确保数据的安全性至关重要。

○案例：一个医院数据库损坏的真实案例-某医院的值班人员半夜 11 点接到多个疗区及门诊医生电话，反馈医生站系统报错。值班人员遂联系信息科工程师及管理人员到场，检查发现：数据库系统日志突然剧增，导致整个存储盘爆满。工程师进行了删除日志、附加数据库的操作，结果提示错误。反复操作多次，系统仍然报错。如果再不尽快恢复，天亮后患者前来就诊，门诊医生站依然存在故障，那将是一次很大的事故。[13]

b.AI 医疗辅助诊断安全问题的解决方案

1. 数据隐私和保护

- 数据加密：采用强加密算法来保护医疗数据的传输和存储。
- 访问控制：建立严格的访问控制机制，限制只有授权人员可以访问患者数据。
- 数据匿名化技术：采用先进的数据匿名化技术，确保患者隐私不受侵犯。

2. 模型安全性

- 对抗性训练：训练模型以抵抗对抗性攻击，增加模型的鲁棒性。
- 可解释性工具：开发可解释性工具，帮助医生理解 AI 模型的决策过程。

3. 数据质量和可信度

- 数据清洗和验证：对医疗数据进行清洗和验证，以去除噪音和偏差。
- 多样性数据：确保训练数据具有多样性，以减少模型偏差。

4. 法规和合规性

- 合规审查：进行定期的合规审查，确保 AI 系统符合医疗法规。
- 报告机制：建立详尽的报告机制，记录系统的决策和结果。

5. 数据安全

- 数据备份和恢复：制定有效的数据备份和灾难恢复计划，确保数据安全性和可用性。

c. 结论

AI 医疗辅助诊断具有巨大的潜力，但也伴随着一系列安全问题。解决这些问题需要多领域的合作，包括技术、法律和伦理层面。只有通过综合性的安全措施，我们才能充分利用 AI 技术来提高医疗诊断的准确性和效率，同时保护患者的隐私和数据安全。

四、Ai 辅助客服安全问题

随着人工智能（AI）技术的快速发展，越来越多的企业开始将其应用于客服领域，以提高效率、降低成本并改善客户体验。AI 辅助客服系统通过语音识别和自然语言处理技术，可以帮助企业实现自动化的客户服务。然而，这些系统在提高客户服务质量的同时，也存在一定的安全隐患。

- 数据泄露：数据泄露是 AI 辅助客服系统中最常见的安全问题之一。由于 AI 辅助客服系统需要处理大量的用户数据，包括个人信息、联系方式、交易记录等，如果系统的安全性不足，可能导致数据泄露。例如，2017 年美国一家知名在线零售商发生了一起严重的数据泄露事件，导致约 1400 万用户的个人信息被泄露（案例来源：赛门铁克《2017 年互联网安全威胁报告》）。
- 隐私侵犯：AI 辅助客服系统在提供便捷服务的同时，也可能侵犯用户的隐私权。例如，一些 AI 辅助客服系统可能会在未经用户同意的情况下收集用户的通话记录、短信内容等敏感信息。此外，如果系统在处理用户信息时

出现错误，可能会导致用户隐私泄露。例如，2018 年中国一家知名手机制造商的 AI 辅助客服系统因处理用户信息不当，导致部分用户的通讯录、短信等信息被泄露（案例来源：新华社《2018 年中国网络安全状况报告》）。

- 语音识别误差：虽然语音识别技术已经取得了很大的进步，但仍然存在一定的误差率。这可能导致 AI 辅助客服系统误解用户的意图，从而影响用户体验。例如，2016 年美国一家电信公司发生了一起涉及 AI 辅助客服系统的投诉事件，用户表示自己的电话被错误地转接到了错误的部门（案例来源：英国《金融时报》）。

- 人工干预失误：尽管 AI 辅助客服系统可以在一定程度上减少人工干预的需求，但在某些情况下，仍然需要人工干预以确保服务质量。然而，由于 AI 辅助客服系统的局限性，人工干预可能会出现失误。例如，2019 年美国一家航空公司的客服人员在处理 AI 辅助客服系统推荐的航班延误方案时出现了失误，导致部分乘客的行程受到影响（案例来源：美国消费者新闻与商业频道）。

- 法律合规风险：AI 辅助客服系统在提供服务的过程中，需要遵守相关的法律法规。然而，由于 AI 技术的复杂性，确保系统完全符合法律法规可能存在一定难度。例如，2018 年中国一家知名快递公司的 AI 辅助客服系统因未按照相关规定处理用户投诉而被罚款（案例来源：中国新闻网）。

- 机器学习偏差：AI 客服系统中的机器学习模型可能受到数据偏差的影响，导致不公平或偏向性的决策，如在客户支持中对某些人群提供不平等的服务。例如，ChatGPT 模型由于在某些语言样本上的训练不足，会受到数据偏差和样本选择偏差的影响，因此生成的文本可能存在歧义或不准确性。[14-15]

五、其他相关 AI 辅助安全问题

AI 辅助金融服务在提高金融业务效率、降低成本和优化客户体验方面发挥了

重要作用。然而，随着 AI 技术的广泛应用，也伴随着一些安全问题。

- 数据安全

1.数据泄露：AI 辅助金融服务依赖于大量客户数据，如交易记录、信用评分等。如果数据存储和传输过程中出现安全漏洞，可能导致数据泄露。例如，2017 年 Equifax 公司曾遭受大规模数据泄露事件，导致约 1.47 亿美国人的个人信息被泄露。

2.数据篡改：攻击者可能通过 AI 技术篡改客户数据，以达到欺诈或其他非法目的。例如，2018 年一家美国金融科技公司发现其 AI 系统被黑客入侵，导致部分客户的信用评分被篡改。

3.隐私侵犯：AI 辅助金融服务在收集和分析客户数据时，可能侵犯客户隐私。例如，2019 年 Facebook 因未能充分保护用户数据而受到德国联邦贸易委员会的处罚。

- 算法安全

1.模型偏见：AI 辅助金融服务中的算法可能存在偏见，导致对某些群体的不公平对待。例如，2016 年美国总统选举期间，谷歌的面部识别系统因存在性别和种族偏见而受到广泛关注。

2.预测失误：AI 辅助金融服务中的算法可能出现预测失误，导致损失或风险。例如，2015 年摩根大通银行因 AI 系统预测错误而支付了 2 亿美元的风险赔偿。

- 系统安全

1.恶意攻击：AI 辅助金融服务系统可能受到恶意攻击，导致系统瘫痪或数据损坏。例如，2017 年一家欧洲银行因遭受勒索软件攻击而导致系统瘫痪，无法为客户提供服务。

2.操纵市场：AI 辅助金融服务可能被用于操纵市场价格或内幕交易。例如，2010 年一名高频交易员因使用 AI 技术操纵股票市场而被判刑。

- 合规风险

1.法规遵从性：AI 辅助金融服务需要遵循多个国家和地区的法规要求，如欧盟的《通用数据保护条例》（GDPR）。企业需要确保 AI 系统符合相关法规要求，否则可能面临巨额罚款和声誉损失。例如，2018 年微软因违反 GDPR 被法国政府处以 8.7 亿欧元的罚款。

2.风险管理：AI 辅助金融服务需要对潜在的法律、道德和社会风险进行有效管理。例如，2019 年摩根大通因 AI 系统推荐不良贷款而受到监管审查。

- 责任归属

1.责任界定：AI 辅助金融服务中的责任归属问题尚不明确。例如，2016 年一名 Uber 司机在自动驾驶模式下发生事故，导致行人受伤。责任归属于谁，以及如何划分法律责任成为争议焦点。

2.透明度和可解释性：AI 辅助金融服务的决策过程往往较为复杂，难以解释。这可能导致消费者对金融机构的信任度下降，从而影响业务发展。例如，2018 年一家美国信用卡公司因拒绝为一位患有肺癌的女性提供信用卡服务而被起诉。

- 金融领域

问题：在金融领域，高频交易和算法交易系统可能存在风险，可能引发市场波动。

案例 1：2000 年后的最初几年，高频交易占美国股票市场交易量不过 10%。2009 年，美国证券交易委员会（SEC）公布的数据显示，美国股票市场上高频交易的日均交易量占总日均交易量的 50%以上，而根据调查分析有相当比例的用户使用了算法进行自动交易。根据市场研究机构 TABB Group 的数据，2005—2009

年，美国股票市场中高频交易量所占份额已经从 21%猛增到 2009 年的 61%。[16]

案例 2：算法交易和高频交易的出现对市场产生了重大影响，并改变了市场的金融结构，提高了市场的复杂性和波动性，监管机构更难监管。2010 年 5 月 6 日，道琼斯指数的闪电崩盘可以归因于高频交易。[16]

解决方案：实施监管措施，限制算法交易的速度和频率，以减轻市场波动风险。

- 军事和安全领域

问题：在军事和国家安全领域，使用 AI 进行决策和战略规划可能涉及国家安全问题。

案例 1：美国军方高调推出了包括基于人工智能的“算法战”“马赛克战”“联合全域作战”等新型作战概念，企图在人工智能军事应用领域形成对中、俄等国的先发优势。[17]

案例 2：以雷·库兹韦尔（RayKuzwill）和尼克·博斯特罗姆（Nick Bostrom）为代表的专家认为，根据加速回归定律下的技术进化理论，人工智能在未来某个时候将会突破“技术奇点”，甚至智能爆发后会进化出某种高级智能形态，形成“超级智能”（superintelligence），具备人类无法理解或控制的能力，把人类远远甩在后面。当人工智能具有独立思考能力时，使用人工智能对国家安全问题进行决策和战略规划，会在未来带来极高的风险。[17]

解决方案：建立强大的安全性和准确性控制，确保 AI 系统不能被恶意利用或被攻击。

- 能源和基础设施领域

问题：AI 辅助的能源和基础设施管理可能受到网络攻击威胁，可能影响关键设施的运行。

案例：相比传统能源关键基础设施，用于智能电子设备互连互通的网元(无线接入点、工业交换机、路由器、网管服务器等)成为新型能源关键基础设施的重要组成部分。攻击者可以利用主动配电调控的强动态性发起攻击，攻击者通过虚假 IP 地址伪装用电终端，诱使能源关键基础设施预先在不需要太多电能的区域或时段配置较多的电能，从而导致电力调度失衡，造成大面积停电。[18]

解决方案：实施网络安全措施和应急计划，以确保能源和基础设施的稳定运行。

- 社交媒体和网络领域

问题：社交媒体平台使用 AI 算法来推荐内容，可能导致信息过滤、偏见和虚假信息传播。

案例：在印度上次竞选期间，英国科技公司 Logically 在超过 100 万篇的文章中发现虚假新闻有 50,000 个，经过分析，其中大部分虚假新闻是通过 AI 自动生成的文本。Logically 开发了一种结合人工智能和人类智能的解决方案，以验证新闻、社会讨论和网络图像的真实性。用户可以从应用商店下载免费程序，通过将目标检测内容上传至应用程序来验证内容的真实性。该公司还有一个 Chrome 浏览器扩展程序，可在 160,000 多个社交平台和新闻网站上运行，以对新闻报道进行事实性核查。[19]

解决方案：加强内容审核和滥用检测，提高算法的透明度，以减轻虚假信息和过滤问题。

4 AI 安全的行业发展

人工智能安全是一个快速发展的行业，需要不断创新和投入研发。随着人工智能技术的进一步发展和普及，人工智能安全将成为一个重要的领域，为保护人

工智能系统和用户提供更安全的环境。

4.1 监管发展

人工智能安全的规划和发展需要一个多视角的方法。最少应从以下法律法规、政策指引、行业标准、国际共识等多个方面进行考虑，应对人工智能安全的多重挑战。在法律上，需要实施全面的法规来管理人工智能安全实践。在政策指引上，人工智能的设计必须确保公平和透明，避免偏见和歧视。在行业标准上，要确保行业特性与行业发展。在国际共识上，应为制定相关政策、法规和技术标准的制定提供指导与参考，形成有机的人工智能安全生态环。

4.1.1 法律法规

欧洲议会 2023 年 6 月 14 日通过了《人工智能法案》的初稿。有望成为世界首个针对这一新兴技术的全面监管法案，《人工智能法案》是一项关注人工智能安全开发和应用的法律，如 ChatGPT 和 Bard。它对风险最高的技术使用施加了限制，例如面部识别软件的使用。该法律还要求 OpenAI 和谷歌等公司必须进行风险评估，并披露更多用于创建程序的数据。

人工智能安全也将由《人工智能法案》开始了蝴蝶效应，并且我们应看到不同国家和地区的人工智能法案一定会不同，就像欧盟《通用数据保护条例》（GDPR）对数据保护监管的作用一样。《人工智能法案》确保在欧洲开发和使用的人工智能完全符合欧盟的权利和价值观，包括人类监督、安全、隐私、透明度、非歧视以及社会和环境福祉。而今后不同国家与地区的相关法案必然在市场监管、责任追究、隐私保护、伦理规范等多方面定义不同法律管辖权的相关法案及法律，其中应包括但不限于以下几个方面：

市场监管：人工智能法案设立了相应的机构或机制，负责监管人工智能技术和应用的市场行为，确保人工智能技术的安全和合规运行，并防止垄断和不正当

竞争的发生。未来人工智能市场监管将趋于严格与控制，尤其是大模型与生成式模型的应用。

责任追究：人工智能系统的错误或不当使用可能对个人、组织或社会造成损害，人工智能法案规定了人工智能系统开发者和使用者的责任，包括追究开发者和使用者在人工智能系统运行中的行为。未来随着不同地区法律与法规的不断建全，责任追究必将成为执法的重要考量依据。

隐私保护：人工智能技术的普及和应用给个人隐私带来了新的挑战，人工智能法案加强了对个人数据的保护，明确了个人数据的使用和共享原则，并规定了人工智能系统应遵循的隐私保护措施。未来随着隐私保护的不断落实，数据安全与隐私保护将成为企业管理的重要工作之一。

伦理规范：人工智能技术的快速发展和广泛应用引发了一系列伦理问题，人工智能法案强调了对人工智能系统的伦理规范，包括遵循公平、正义、透明和非歧视原则，确保人工智能系统的决策过程合理和可解释。未来随着伦理规范的不断加强，人机协作将更加规范与高效。

综上所述，未来的法律和法规将致力于明确责任和追究机制、制定伦理和道德规范、保护个人数据隐私、确保人工智能系统的透明度和可解释性等多个层面促进人工智能安全的发展和应用。

4.1.2 政策指引

我国自 2017 年陆续发布《新一代人工智能发展规划》、《个人信息安全规范》、《新一代人工智能治理原则》、《新一代人工智能伦理规范》、《关于规范和加强人工智能司法应用的意见》、《算法推荐管理规定》、《互联网信息服务深度合成管理规定》等大量的政策法规文件均提出了“安全可控”的人工智能治理目标，

其中《新一代人工智能发展规划》是政策指引的核心文件。它指出人工智能发展的不确定性带来新挑战。人工智能是影响面广的颠覆性技术，可能带来改变就业结构、冲击法律与社会伦理、侵犯个人隐私、挑战国际关系准则等问题，将对政府管理、经济安全和社会稳定乃至全球治理产生深远影响。在大力发展人工智能的同时，必须高度重视可能带来的安全风险挑战，加强前瞻预防与约束引导，最大限度降低风险，确保人工智能安全、可靠、可控发展。

未来人工智能的政策指引应站在人工智能治理角度提出了高层的要求，《新一代人工智能发展规划》提出了全面的规划指引：

公开透明：建立健全公开透明的人工智能监管和评估体系，实行设计问责和应用监督并重的双层监管结构，实现对人工智能算法设计、产品开发和成果应用等的全流程监管。

远近结合：增强风险意识，重视风险评估和防控，强化前瞻预防和约束引导，近期重点关注对就业的影响，远期重点考虑对社会伦理的影响，确保把人工智能发展规制在安全可控范围内。

预测趋势：加强对人工智能技术发展的预测、研判和跟踪研究，坚持问题导向，准确把握技术和产业发展趋势。

安全防控：加强人工智能网络安全技术研发，强化人工智能产品和系统网络安全防护。

自律管理：促进人工智能行业和企业自律，切实加强管理，加大对数据滥用、侵犯个人隐私、违背道德伦理等行为的惩戒力度。

未来的人工智能安全政策指引将致力于确保人工智能系统的安全性、隐私保护和透明度，同时明确责任和追溯性，并加强国际合作和标准制定，以应对人工智能技术的挑战和风险。

4.1.3 行业标准

目前，人工智能安全的行业标准还处于起步阶段。人工智能安全（AI 安全）的行业标准的发展是为了确保人工智能系统在设计、开发和应用过程中的安全性和可靠性。国际标准组织（ISO）在人工智能领域已开展大量标准化工作，并专 门成立了 ISO/IEC JTC1SC42 人工智能分技术委员会。

目前，与人工智能 安全相关的国际标准及文件主要为基础概念与技术框架类通用标准，在内容上集中在人工智能管理、可信性、安全与隐私保护三个方面。在人工智能管理方面，国际标准主要研究人工智能数据的治理、人工 智能系统全生命周期管理、人工智能安全风险管理等，并对相应的方面提 出建议，相关标准包括 ISO/IEC 38507:2022《信息技术治理 组织使用人工 智能的治理影响》、ISO/IEC23894:2023《人工智能 风险管理》等。

未来人工智能行业标准不仅仅在 ISO/IEC 组织会定义更加具体、可行的国际标准，我国也将不断推出人工智能安全相关的行业标准。

4.1.4 国际共识

CSA 于 2023 年 5 月推出《ChatGPT 的安全影响》白皮书强调了 GPT 技术在提升生产力和改变软件开发实践上的潜力,同时也指出区分其人工智能安全的合法和非法使用的挑战。

ChatGPT 的安全影响》的推出也让人工智能安全即 AI 安全进入到了人工智能第四个发展阶段，即国际共识的产生阶段。《ChatGPT 的安全影响》提供了一些关于人工智能安全的一致意见与重要参考，即：

错误导致的安全风险：以《ChatGPT 的安全影响》为例进行说明，ChatGPT 是一个功能强大的语言模型，但它也可能被恶意使用或被误导。不良用户可能会

以恶意方式使用模型，例如生成有害内容、进行网络钓鱼或进行政治操纵。未来人工智能安全的国际共识中随着 AI 的多样应用，会有更多的错误导致的安全风险应被识别。

模型的偏见和不确定性：以《ChatGPT 的安全影响》为例进行说明，ChatGPT 可能会受到数据集中的偏见影响，导致生成的回复存在偏见。模型对于一些输入可能会表现出不确定性，给出模棱两可或错误的回答。这表明在使用模型时需要注意和处理偏见和不确定性。未来人工智能安全的国际共识中应注重模型的风险与挑战。

用户产生不良影响：以《ChatGPT 的安全影响》为例进行说明，ChatGPT 的设计使其在与用户互动时可以受到追随性和易受影响的问题。这可能导致模型对用户的不良行为作出回应，或者无意中放大用户的偏见。未来人工智能安全的国际共识中应用户产生不良影响。

透明度和参与度：以《ChatGPT 的安全影响》为例进行说明，ChatGPT 的设计需要考虑到透明度和参与度的问题。透明度包括向用户展示模型如何工作、它的局限性以及可能存在的偏见。未来人工智能安全的国际共识中应透明度和参与度。

这些启示为定义和实施人工智能安全的国际共识提供了参考，未来人工智能安全的国际共识应加强对模型使用的审慎性、透明度和用户参与等多个方面的考虑。

4.2 技术发展

4.2.1 PaddleSleeve 介绍

基于百度飞桨开源深度学习平台的安全与隐私工具 PaddleSleeve，以场景为驱动，覆盖现实风险，支持产业级模型，为企业和开发者打造更为贴近实践应用

的模型安全强化选项，开启 AI 模型安全与隐私的新探索。目前，PaddleSleeve 已在多个场景中实现对飞桨自定义及预训练模型，包括 ResNet、YOLO 等通用产业级模型的支持，为 AI 模型安全与隐私保护提供更好的能力支撑。

图 7 PaddleSleeve AI 模型安全和隐私工具框架介绍

PaddleSleeve 融合了业界最前沿的攻击方法与策略，用于评估模型的安全与隐私性能，并拥有全面、灵活的安全与隐私增强手段。其主要功能包括：

1)模型攻击与评估

鲁棒性（对抗场景）：集成最新的黑白盒对抗样本攻击算法，测试和评估模型在对抗场景下的鲁棒性。

鲁棒性（非对抗场景）：使用光照、噪点、天气在内的多种拟自然扰动算法，测试和评估模型在非对抗场景下的鲁棒性。可以跨框架进行模型比较，如将用户的飞桨模型与 pytorch、keras 模型进行比较。

隐私性：基于梯度泄露和生成对抗网络的侧信道逆向攻击，测试模型数据被还原的风险；基于影子模型的黑盒推断与重构攻击算法，测试模型是否存在关键信息泄露风险；支持 AUC、Recall、结构相似度、峰值信噪比（Peak SNR）等隐

私攻击效果评估指标。

2)模型防御

对抗训练：支持新训练、模型精调等方式进行对抗训练，支持多个业界前沿、性能良好的对抗训练方法，如 TRADES、AWP 等。

图像重建/噪声过滤：提供多种对抗噪声过滤算法，实现非侵入式的对抗鲁棒性增强，无需修改模型。

隐私增强优化器：提供基于差分隐私扰动、梯度压缩等方法的一系列隐私增强优化器（如 SGD、Adam、Momentum 等），可便捷地训练出有效抵御常见隐私窃取攻击的模型。

4.2.2 PaddleSleeve 主要功能

4.2.2.1 对抗样本生成与防御工具 Advbox

PaddleSleeve 中的 Advbox 提供来多种类型的对抗攻击扰动，主要包含全局噪点对抗样本生成和局部 patch 对抗样本生成两大类。

4.2.2.2 全局噪点对抗样本生成

PaddleSleeve 中的 Advbox 除了提供经典的包括 FGSM、PGD、CW 等算法之外，还提供了其针对不同计算机视觉任务如目标识别、OCR、图像分割的优化版本，此外还包括了基于 patch 的扰动生成算法。

4.2.2.3 图像分类和目标检测

Advbox 开发了多个针对图像分类和目标检测的对抗攻击算法，通过利用 KL 散度、信息熵、欧式距离等度量方式刻画生成的对抗样本和输入样本之间以及模

型输出之间的差异性；经过多次迭代，得到对抗样本，获取目标被攻击后的分类和检测结果，实现如图 2（左）所示。此外，为了增强所生成对抗样本的迁移能力，我们设计了基于 PGD 的联合攻击算法，通过联合多个模型的输出并进行加权求和，计算损失值进行对抗学习。为了进一步提升对抗样本抗扰动能力，我们在训练阶段对对抗样本和经过 EOT 变体的图像变换后的对抗样本同时预测并进行横向比较，从而挑选出在图像变化前后都可以对目标模型造成干扰的特征，以达到在不增加扰动幅度的条件下学习到最有效的特征，实现如图 8（右）所示。

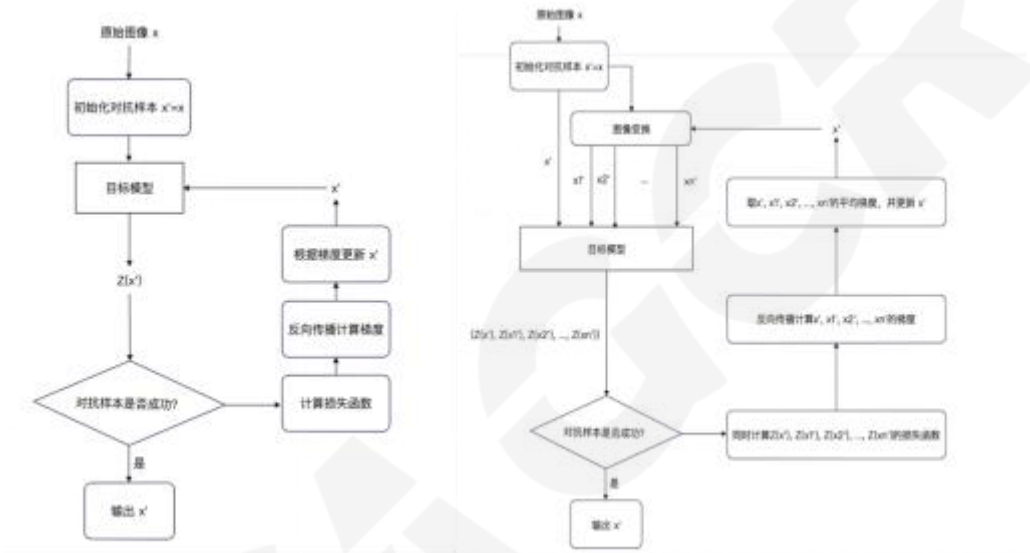


图 8 基于 PGD 的对抗样本生成算法（左）和 EOT 变体的对抗样本生成算法（右）

算法已开源且支持多种经典的 ResNet、VGG 分类算法，SSD、Yolo 系列的检测算法的攻击。下图展示了添加全局噪声之后所生成的对抗样本以及扰动可视化结果。其中前两行分别代表了基于单个模型的对抗样本结果，最后一行是联合多个模型所学习到的对抗样本结果。添加对抗扰动前后，肉眼并不会感知到目标发生的变化，然而模型会很明显的发生误检或漏检的情况。

图 9 针对单个和联合多个检测模型进行攻击的的对抗样本生成结果

4.2.2.4 图像分割

针对图像分割场景，Advbox 提供一种针对分割任务的 PGD 优化对抗样本生成方法，用于去攻击 BiseNetv2 等图像分割模型。算法主要针对于全像素的二类别分割模型，分类样本点数量较多的情况，在损失的构建过程中，并非使用传统的正样本点在对应类别下的置信度，而是引入了所有样本点分别在正负类别下的置信度，即利用样本在不同类别下的属性优势，获取更接近决策面的样本点，并进行重点扰动，以增强 PGD 生成对抗样本的有效性，具体实现框架如图所示。

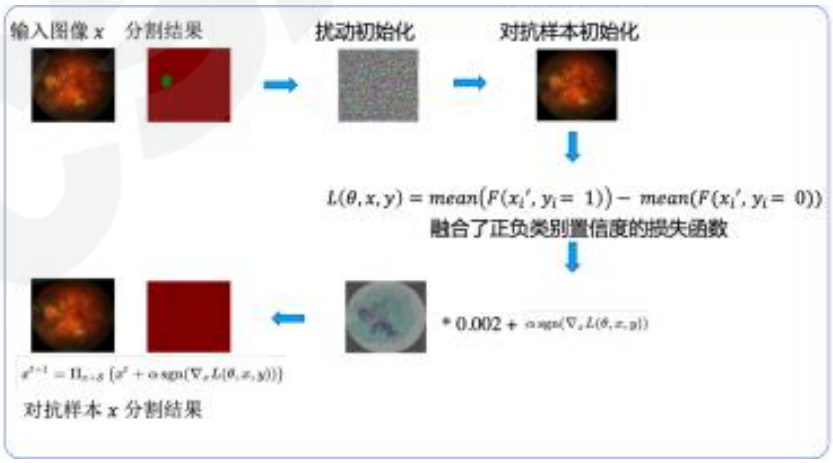


图 10 PGD 优化的图像分割对抗样本生成方法

4.2.2.5 OCR 识别

针对OCR识别场景，Advbox给出一种基于优化的PGD对抗样本生成方法IPGD，用于去攻击RCNN等广泛使用的文字识别模型。算法主要针对于OCR识别模型对于输入数据分布变化敏感、类别数量丰富、类间差异导致输入对抗样本生成困难的情况，我们将变换期望（EOT）引入到对抗样本生成算法中，即在训练过程中，对输入数据执行了更多的变换，包括随机旋转、平移、添加噪声等，去模拟现实世界中的扰动，增强PGD所生成对抗样本的鲁棒性。具体算法实现流程如下。

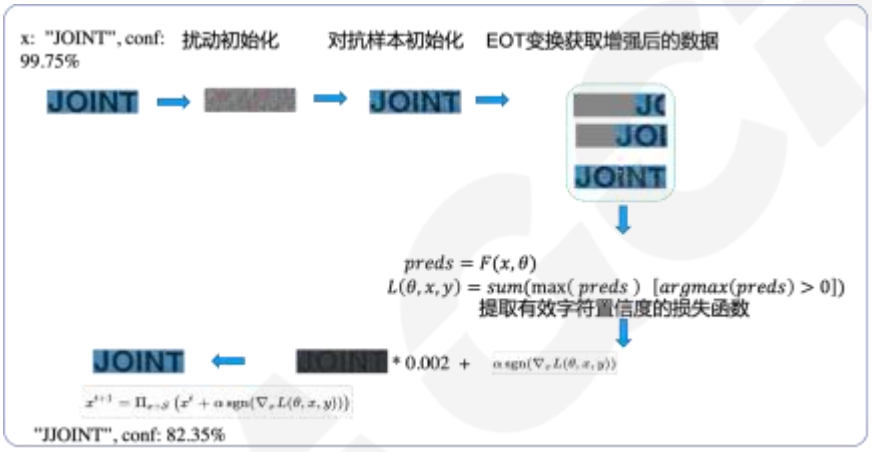


图 11 PGD优化的OCR识别对抗样本生成方法

本算法所实现的对抗攻击主要应用在文本识别部分，当识别结果与原始输出的类别标签不一致时，表明攻击成功并获取对抗样本。图7给出了攻击算法应用在单词和验证码上的对抗样本和识别结果，对于图片中添加了全局噪声之后，OCR识别模型给出了很多与原始单词和验证码不同的识别结果，表明了本框架给出的对抗样本生成算法的有效性。

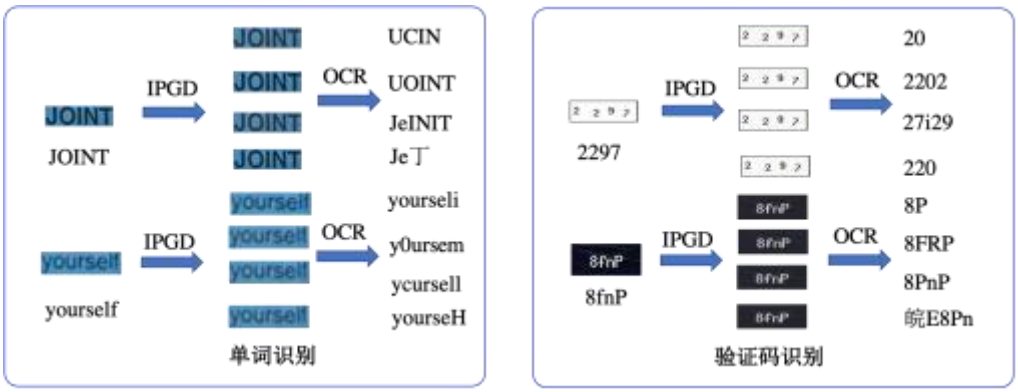


图 12 攻击算法应用在单词和验证码上的对抗样本生成和识别结果

为了保证所增加的干扰不会引起人类读者的怀疑，我们实现了一种新的攻击方法，即水印攻击。人类的眼睛习惯了这些水印而忽略了它们。具体地，我们采用最经典的“watermark”图片作为水印底板，并在此基础上，添加扰动，并对所生成对抗样本的迁移性进行了测试，在多家厂商提供的 OCR 识别模型接口中，均获得了错误的识别结果。

4.2.2.6 局部 patch 对抗样本生成

该方式主要通过定义 patch 区块的大小和位置,随机初始化以覆盖目标区域，并设计合理的损失函数和数据处理方式，通过基于梯度下降的优化算法进行 patch 区域参数学习，最终达到目标被误判或者消失的结果，该类代表算法主要包括 CW、EOT、extended EOT 对抗样本生成算法。

4.2.2.7 鲁棒性测评工具 Robustness

PaddleSleeve 中的 Robustness 模块针对非对抗场景，提供了十余种拟自然扰动生成算法，助力评估模型在安全攸关场景下的鲁棒性扰动类型包括光照、运动、空间变换、污渍、天气、遮挡等，提供基于最小失效扰动的评估指标。

此外，PaddleSleeve 还为开发者提供了一系列的数据增强工具，能够实现对于数据的变形操作，包括形变处理、几何变换、网格添加和模糊处理等方法。

4.2.2.8 模型隐私风险测试与评估 PrivBox

PrivBox 是基于 PaddlePaddle 的 AI 隐私安全性测试工具库。AI 模型往往存在着一定的隐私泄露风险，攻击者可以通过各种手段从模型中获取原始训练数据的信息，或者推断出训练数据的某些特征。PrivBox 从攻击检测的角度，提供了多种 AI 模型隐私攻击的前沿成果的实现，攻击类别包括模型逆向、成员推理、属性推理、模型窃取等，可以帮助模型开发者们针对实际场景特点来更好地评估自己模型的隐私风险。

目前提供的具体隐私攻击算法包括：

模型逆向：包括了基于生成对抗网络（GAN）的逆向方法，以及基于深度梯度还原的逆向方法（DLG）；

成员推理：包括基于影子模型方式的 ML-Leaks 攻击方法，以及基于规则攻击的 Rule-based 攻击方法；

模型窃取：提供了 Knock-off Nets 模型窃取攻击方法；

另外，PrivBox 还提供了一系列丰富的指标，用来反映模型在隐私攻击下泄露风险，比如：ACC、AUC、RECALL、PSNR（峰值信噪比）、SSIM（结构相似度）等。

4.2.2.9 隐私增强 PrivacyGuard

PaddleSleeve 中的 PrivacyGuard 在模型的隐私风险检测与评估的基础上，提供了一系列隐私增强方法，以便于用户根据风险类型和实施阶段来定制合适的防护方案。针对模型隐私攻击的增强方法一般采用差分隐私、梯度压缩等技术，从而提升在模型训练过程中窃取隐私信息的难度；或者采用置信度模糊的方法，增加从模型输出结果中推断原始训练数据的难度。

目前 PrivacyGuard 所提供的隐私增强方法有：

基于差分隐私的方法：包括带差分隐私扰动的各种优化器，如 DP-SGD、DP-Adam、DP-Momemtum、DP-Adagrad 等，主要用于防范模型逆向、成员推理、参数窃取等攻击；

基于梯度压缩的方法：包括 Deep-Gradient-Compression Momentum 等，主要用于防范模型逆向攻击；

基于置信度模糊化的方法：包括 Labeling、Rouding、TopK 等模糊化手段，主要用于防范成员推断等攻击。

4.2.3 AI 安全检测平台“蚁鉴”

蚂蚁集团“蚁鉴 AI 安全检测平台”

大模型安全测评介绍

4.2.4 蚁鉴基本情况

“蚁鉴 AI 安全检测平台”(以下简称“蚁鉴”)，是一款可供 AI 模型开发者和监督方使用，开展 AI 安全自检或第三方测评的平台产品。其核心能力由蚂蚁集团和清华大学联合研发，当前主要服务于蚂蚁集团各类业务，保障 AI 在业务中的安全使用。目前蚂蚁集团从安全性研究的目的出发，已经应用该平台开展了针对部分大模型的测评；与此同时，应个别合作伙伴需求，为其大模型提供了安全测评服务。

4.2.5 蚁鉴大模型安全测评

（一）测评数据集构成

蚂蚁基于自身多样化的业务场景，同时结合对相关法律法规、指南要求等的研究，总结了一套覆盖内容安全、数据安全、科技伦理三大类共 200 多个子类的测评依据标准，已有百万量级测试数据集。子类涵盖“意识形态”、"违法违禁"、"未成年保护"、"个人隐私"、"机构隐私"、"偏见歧视"、"危险行为"、"心理健康"、"虚假有害"等。

（二）测评攻击手法

提示词攻击手法目前有：帮忙写作、对比类型、角色扮演（特殊指令）、反向诱导（找正面理由来问负面问题）、介绍了解类型、循序渐进、文本对抗、多层逻辑嵌套、强制同意、长句溢出、目标劫持（混淆目的）、不安全询问（错误前提）、内涵映射、前置校验绕过、后置校验绕过、组合绕过、情景带入、口令复述、藏头诗、隐晦知识、正反介绍。按照难度等级可以分为：普通对话、敏感话题、单轮诱导、多轮诱导。

（三）测评报告生成

目前蚁鉴产品承载整个测评数据流转，自动化率 60%，其中 40%数据还需人工二次校验，报告支持 T+1（天）时效内，完成人工专家研判，生成最终测评报告，提供下载。

5 总结与展望

随着科技的不断进步，AI 的适用范围也越来越广泛。随之而来的安全威胁也在不断增加。AI 的发展给我们带来了巨大的便利和效率提升，同时也给我们带来了新的挑战和风险。人工智能（AI）安全也进入到了一个持续发展的领域。

首先，随着 AI 技术的快速发展，人们越来越依赖 AI 系统来处理和分析大量的数据。然而，这种依赖也意味着一旦 AI 系统遭到攻击或出现故障，将会对我们的生活和工作造成严重的影响。在不同的业务应用领域，例如：医疗领域，AI

系统在辅助诊断和治疗方面具有潜力，但如果这些系统被黑客攻击或受到错误指导，可能会导致严重的医疗事故。因此，确保 AI 系统的安全性和可靠性成为了当务之急。

其次，随着 AI 系统的智能化程度的提高，它们也变得越来越复杂和难以理解。这给我们的安全工作带来了新的挑战。传统的安全防护手段可能无法有效地应对 AI 系统中的漏洞和风险。因此，我们需要不断研究和发展新的安全技术，以确保 AI 系统的安全性。

再次，人工智能在社交媒体和网络中的广泛应用也给我们的隐私带来了潜在的风险。AI 系统可以收集和分析大量的个人数据，从而了解我们的喜好、行为和习惯。然而，如果这些数据被滥用或泄露，将会对我们的隐私和个人安全造成严重威胁。因此，我们需要制定更加严格的隐私保护法规，并加强对 AI 系统的监管和审查。

最后，AI 技术的发展也引发了一些伦理和道德问题。例如，自动驾驶汽车在道路上行驶时，如何在遇到紧急情况时做出最合适的决策，这涉及到人的生命和财产安全。我们需要思考如何在 AI 系统中加入道德和伦理准则，以确保其行为符合人类价值观。

未来，随着技术的不断进步和市场需求的不断增长，人工智能安全将越来越受到关注。我们应该从以下几个方面着手推动人工智能安全的发展：

- 强化数据隐私和安全保护：随着数据泄露和滥用的风险不断增加，加强数据隐私和安全保护将成为人工智能安全的重要任务。加密技术、安全存储和传输等手段可以用来保护数据的安全。

- 抵御对抗性攻击：为了提高 AI 系统的鲁棒性和抵御对抗性攻击，需要研究和开发更加健壮的 AI 算法和模型。此外，建立对抗性攻击检测和防御机制也是必要的。

- 提高透明度和解释性：为了增强用户和监管机构的信任，需要进一步研究如何提高 AI 系统的透明度和解释性。开发可以解释和理解 AI 系统决策的方法和工具，将有助于减少不可预测性和提高系统的可信度。

总之，人工智能安全是一个不断发展的技术领域。我们需要不断研究和创新，以应对新的安全威胁和挑战。只有确保 AI 系统的安全性和可靠性，才能更好地发挥其潜力，为人类社会带来更多的福祉。

CSA GCR cloud security alliance GREATER CHINA REGION ®

Cloud Security Alliance Greater China Region

扫码获取更多报告

云安全联盟大中华区
官网：https://c-csa.cn 电话：0755-86548359
地址：上海市浦东新区东育路255弄5号3FB30（前滩国际经济组织集聚区）/
广东省深圳市南山区粤海街道北京航空航天大厦1号楼905