蚂蚁集团 ANT GROUP | 支付宝 ALIPAY

人工智能安全白皮书（2020）

浙江大学-蚂蚁集团金融科技研究中心
数据安全与隐私保护实验室

2020年12月

人工智能安全白皮书

目录

第一章　引言

人工智能（Artificial Intelligence，AI）战略是国家战略。习近平总书记在十九届中央政治局第九次集体学习时指出“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”。2020 年 11 月，国务院办公厅印发了关于切实解决老年人运用智能技术困难实施方案的通知，提到鼓励在就医场景中应用人脸识别等技术。

人工智能技术的崛起依托于三个关键要素：1）深度学习模型在机器学习任务中取得的突破性进展；2）日趋成熟的大数据技术带来的海量数据积累；3）开源学习框架以及计算力提高带来的软硬件基础设施发展。我们在白皮书中将这三个因素简称为 **AI 模型**、**AI 数据**以及 **AI 承载系统**。在这三个要素的驱动下，AI 技术已经成功应用于生物核身、自动驾驶、图像识别、语音识别等多种场景中，加速了传统行业的智能化变革。例如：在自动驾驶场景中，AI 技术对车辆、行人、路牌和路标等环境信息进行感知来辅助驾驶系统进行决策，并结合云端交通信息自动完成路线规划，使得自动驾驶汽车可以安全、高效地完成驾驶任务；在用户网购场景中，AI 技术通过对用户基本属性、浏览点击信息以及商品属性等数据进行挖掘分析，预测用户行为偏好，从而提供更加个性化的推荐服务。总而言之，随着对这三个因素的探索持续深入，AI 技术不仅在多个经典机器学习任务中取得了突破性进展，还广泛应用于真实世界中的各类场景。

人工智能推动社会经济各个领域从数字化、信息化向智能化发展的同时，也面临着严重的安全性威胁。面对人工智能安全性威胁，学术界和工业界“抓住机遇，迎难而上”，对人工智能安全技术（AI 安全）进行了前瞻性研究与布局。研究发现，这些安全性威胁极大程度上破坏了人工智能技术良性发展的生态。

这些威胁一方面会严重损害 AI 技术的功能性，例如攻击者可以通过恶意篡训练数据、污染 AI 模型的训练过程，来破坏 AI 模型功能性 [1]。攻击者甚至可以对训练数据嵌入特定的“后门”（Backdoor），在不影响 AI 模型在正常数据集判别性能的情况下，操纵其对携带“后门”数据的判断结果。例如：在图像分类任务中，攻击者可以在图片的角落添加一块特殊图案作为“后门”。这样训练后的模型在接收到含有“后门”图案的图片后，就会被操纵做出指定的判断 [2]。研究者还发现在输入数据上添加少量精心构造的人类无法识别的“扰动”，可以使 AI 模型输出错误的预测

结果 [3]。这种添加扰动的输入数据通常被称为对抗样本（Adversarial Example）。在许多安全相关的应用场景中，对抗样本攻击会引起严重的安全隐患。以自动驾驶为例，攻击者可以在路牌上粘贴对抗样本扰动图案，使得自动驾驶系统错误地将“停止”路牌识别为“限速”路牌 [4]。这类攻击可以成功地欺骗特斯拉等自动驾驶车辆中的路标识别系统，使其作出错误的驾驶决策判断 [5]，导致严重的交通事故。另一方面，AI 技术还面临着严峻的隐私泄露威胁。研究者发现 AI 技术在使用过程中产生的计算信息可能会造成隐私数据泄露，例如攻击者可以在不接触隐私数据的情况下利用模型输出结果、模型梯度更新等信息来间接获取用户隐私数据。在实际应用中，这类信息窃取威胁会导致严重的隐私泄露。例如：生物核身识别模型返回的结果向量可以被用于训练生成模型，从而恢复如用户头像等训练数据中的敏感信息 [6]。攻击者甚至还可以通过输出结果窃取 AI 模型的参数 [7, 8]，对模型拥有者造成严重的经济损害。本白皮书将在第三章系统性地归纳和总结 AI 模型、AI 数据与 AI 承载系统面临的安全威胁，并根据不同的应用场景与攻击者的能力假设，分析近年来相关研究的优缺点，并探讨现有攻击技术的进一步发展趋势。

为了应对 AI 技术的安全与隐私泄露威胁，学术界与工业界深入分析攻击原理，并根据不同的攻击原理提出一系列对应的防御技术。这些防御技术覆盖了数据收集、模型训练、模型测试以及系统部署等 AI 应用的生命周期，充分考虑了每个阶段可能引发的安全与隐私泄露威胁，详细分析了现有攻击方法的原理、攻击实施的过程以及产生的影响，并最终提出对应的防御技术。例如：为了防止攻击者在数据收集阶段污染训练数据并操纵模型训练参数，研究者分析了训练数据毒化对模型产生的影响，随后提出了利用聚类模型激活神经元来区分毒化和干净的数据的防御方法 [9]；为了防止已经训练好的 AI 模型被嵌入攻击“后门”，研究者分析了模型中存在“后门”攻击的潜在特征，随后提出了模型剪枝/微调等方法来消除模型中存在的“后门” [10]；为了防止攻击者在测试阶段发起的对抗样本攻击，研究者提出使用 JPEG 压缩、滤波操作、图像模糊处理等方法对输入数据进行预处理，从而降低对抗性扰动带来的影响 [11]。此外，为了防止 AI 模型在训练/测试阶段泄露模型的关键参数，研究者通过对模型结构的适当调整，降低模型过拟合度，从而减少模型泄露的参数信息。尽管上述研究为 AI 模型提供了有效的防御机制，但会不可避免地降低 AI 技术在应用中的判断准确率和执行效率。除了从技术层面防范 AI 安全威胁之外，越来越多的国家和地区推出了数据安全法律法规来保护用户的隐私数据，例如：欧盟的《通用数据保护规范（GDPR）》、美国的《NIST SP 800-122》以及中国的《中华人民共和国数据安全法（草案）》，这些加速落地的法律法规也为 AI 模型拥有者带来了相应的合规风险。本白皮书将在第四章详细回顾并总结针对经典 AI 攻击技术在不同场景下的主流防御策略，并探讨针对新型攻击手段的防御机制在未来的发展方向。

综上所述，AI 技术所面临的多种安全威胁将会对用户隐私数据造成泄露，并在实际应用场景中对用户的生命与财产带来损失的风险。为了应对 AI 技术所面临的安

全与隐私威胁，本白皮书系统性地总结了学术界与工业界对 AI 安全与隐私保护技术的相关研究成果。在白皮书中，我们聚焦于 AI 技术中模型、数据与承载系统的安全问题。我们将首先详细介绍 AI 模型、数据与承载系统面临的安全威胁，然后逐一介绍针对这些威胁的防御技术，最后提出 AI 应用的一站式安全解决方案。

第二章　AI 技术与安全模型

在本章，我们首先定义 AI 技术所涵盖的技术范围和组成部分；然后，我们针对 AI 技术所包含的技术特点，提出 AI 技术安全的具体定义和安全属性；最后，我们探讨 AI 安全威胁的分类方法。

人工智能是一种通过预先设计好的理论模型模拟人类感知、学习和决策过程的技术。完整的 AI 技术涉及到 AI 模型、训练模型的数据以及运行模型的计算机系统，AI 技术在应用过程中依赖于模型、数据以及承载系统的共同作用。

AI 模型　模型是 AI 技术的核心，用于实现 AI 技术的预测、识别等功能，也是 AI 技术不同于其它计算机技术的地方。AI 模型具有数据驱动、自主学习的特点，负责实现机器学习理论和对应算法，能够自动分析输入数据的规律和特征，根据训练反馈自主优化模型参数，最终实现预测输入样本的功能。AI 模型通常结合数据挖掘、深度神经网络、数值优化等算法层面的技术来实现其主要功能。以手写数字分类任务为例，AI 模型需要判断输入图像是 0-9 中的哪个数字。为了学习手写数字分类模型，研究者构建训练数据集（例如：MNIST 数据集）$\{x_i, y_i\}$，$i = 1, 2, ..., N$，其中 x_i, y_i 代表某张图像与其对应的数字。模型可以选取卷积神经网络 $y = f_\theta(x)$，其中 θ 为卷积神经网络的参数。在训练过程中，AI 模型使用优化算法不断调整卷积神经网络参数，使模型在训练集上的输出预测结果尽可能接近正确的分类结果。

AI 数据　数据是 AI 技术的核心驱动力，是 AI 模型取得出色性能的重要支撑。AI 模型需要根据种类多样的训练数据，自动学习数据特征，对模型进行优化调整。海量的高质量数据是 AI 模型学习数据特征，取得数据内在联系的基本要求和重要保障。尽管 AI 技术所使用的算法大多在 20 年前就已经被提出来了，但是直到近些年来，随着互联网的成熟、大规模数据的收集和大数据处理技术的提升才得到了迅猛的发展。大规模数据是 AI 技术发展的重要支撑，具有以下几个特点：（1）数据体量大，AI 模型主要学习知识和经验，而这些知识和经验来源于数据，然而单个数据价值密度较低，大体量的数据有助于模型全面学习隐含的高价值特征和规律；（2）数据多样性强，从各种各样类型的海量数据中，模型可以学习到多样的特征，从而增强模型的鲁棒性与泛化能力。

AI 承载系统 应用系统是 AI 技术的根基，AI 技术从模型构建到投入使用所需要的全部计算机基础功能都属于这一部分。一般的 AI 应用部署的流程大致如下：收集应用所需要的大规模数据，使用相关人工智能算法训练模型，将训练完成的模型部署到应用设备上。AI 承载系统为 AI 技术提供重要的运行环境，例如：储存大规模数据需要可靠的数据库技术、训练大型 AI 模型需要巨大的计算机算力、模型算法的具体实现需要 AI 软件框架和第三方工具库提供稳定的接口，数据收集与多方信息交互需要成熟稳定的互联网通信技术。目前构建 AI 应用常使用的主流框架有 Tensorflow、PyTorch 等，框架高效实现了 AI 模型运行中所需要的各种操作，例如：卷积、池化以及优化等。这些框架提供了 AI 技术执行接口供研发人员调用，使其能够通过调用接口快速搭建自定义的 AI 模型，从而不需要花费太多精力关注底层的实现细节，简化了 AI 应用的开发难度，使开发人员能够更深入地关注业务逻辑与创新方法。这些优点使得 AI 技术快速发展，极大地促进了 AI 应用的落地和普及。

2.1 安全模型

学术界与工业界的研究工作表明 AI 技术在应用过程中存在不可估量的安全威胁，这些威胁可能会导致严重的生命和财产损失。投毒攻击 [1] 毒害 AI 模型，使得 AI 模型的决策过程受攻击者控制；对抗样本攻击 [3] 导致模型在攻击者的恶意扰动下输出攻击者指定的错误预测；模型窃取攻击 [8] 导致模型的参数信息泄漏。此外，模型逆向工程 [6]、成员推断攻击 [12]、后门攻击 [13]、伪造攻击 [14] 以及软件框架漏洞 [15] 等多种安全威胁都会导致严重的后果。这些潜在的威胁使模型违背了 AI 安全的基本要求。在本小节中，我们立足于 AI 技术在应用中面临的威胁，借鉴传统信息安全与网络空间安全的标准规范，讨论适用于 AI 技术的安全模型。

AI 技术的崛起不仅依赖于以深度学习为代表的建模技术的突破，更加依赖于大数据技术与 AI 开源系统的不断成熟。因此，我们在定义 AI 安全模型的时候，需要系统性地考虑 AI 模型、AI 数据以及 AI 承载系统这三者对安全性的要求。在 AI 模型层面，AI 安全性要求模型能够按照开发人员的设计准确、高效地执行，同时保留应用功能的完整性，保持模型输出的准确性，以及面对复杂的应用场景和恶意样本的场景中具有较强鲁棒性；在 AI 数据层面，要求数据不会被未授权的人员窃取和使用，同时在 AI 技术的生命周期中产生的信息不会泄露个人隐私数据；在 AI 承载系统层面，要求承载 AI 技术的各个组成部分能够满足计算机安全的基本要素，包括物理设备、操作系统、软件框架和计算机网络等。综合考虑 AI 技术在模型、数据、承载系统上对安全性的要求，我们用保密性、完整性、鲁棒性、隐私性定义 AI 技术的安全模型，如下：

- **保密性 *(Confidentiality)* 要求 *AI* 技术生命周期内所涉及的数据与模型信息不会泄露给未授权用户。**

- **完整性 *(Integrity)* 要求 *AI* 技术在生命周期中，算法模型、数据、基础设施和产品不被恶意植入、篡改、替换和伪造。**

- **鲁棒性 *(Robustness)* 要求 *AI* 技术在面对多变复杂的实际应用场景的时候具有较强的稳定性，同时能够抵御复杂的环境条件和非正常的恶意干扰。例如：自动驾驶系统在面对复杂路况时不会产生意外行为，在不同光照和清晰度等环境因素下仍可获得稳定结果。**

- **隐私性 *(Privacy)* 要求 *AI* 技术在正常构建使用的过程中，能够保护数据主体的数据隐私。与保密性有所区别的是，隐私性是 *AI* 模型需要特别考虑的属性，是指在数据原始信息没有发生直接泄露的情况下，*AI* 模型计算产生的信息不会间接暴露用户数据。**

2.2　AI 安全问题分类

我们在本小节讨论 AI 技术在应用过程中存在的安全威胁的分类方法，并且分析了常见的安全威胁具体违背了安全模型的哪些安全性要求。总体来说，我们根据 AI 技术涉及的三方面：模型、数据、承载系统，将 AI 安全威胁分为三个大类别，即 AI 模型安全、AI 数据安全与 AI 承载系统安全。

- **AI 模型安全问题** AI 模型安全是指 AI 模型面临的所有安全威胁，包括 AI 模型在训练与运行阶段遭受到来自攻击者的功能破坏威胁，以及由于 AI 模型自身鲁棒性欠缺所引起的安全威胁。我们进一步将 AI 模型安全分为三个子类，分别为：1）训练完整性威胁，攻击者通过对训练数据进行修改，对模型注入隐藏的恶意行为。训练完整性威胁破坏了 AI 模型的完整性，该威胁主要包括传统投毒攻击和后门攻击；2）测试完整性威胁，攻击者通过对输入的测试样本进行恶意修改，从而达到欺骗 AI 模型的目的，测试完整性威胁主要为对抗样本攻击；3）鲁棒性欠缺威胁，该问题并非来自于恶意攻击，而是来源于 AI 模型结构复杂、缺乏可解释性，在面对复杂的现实场景时可能会产生不可预计的输出。上述安全隐患如果解决不当，将很难保证 AI 模型自身行为的安全可靠，阻碍 AI 技术在实际应用场景中的推广落地。我们将在3.1小节中具体介绍这些安全威胁。

- **AI 数据安全问题** 数据是 AI 技术的核心驱动力，主要包括模型的参数数据和训练数据。数据安全问题是指 AI 技术所使用的训练、测试数据和模型参数数据被攻击者窃取。这些数据是模型拥有者花费大量的时间和财力收集得到的，涉及用户隐私信息，因此具有巨大的价值。一旦这些数据泄露，将会侵犯用户的个人隐私，造成巨大的经济利益损失。针对 AI 技术使用的数据，攻击者可

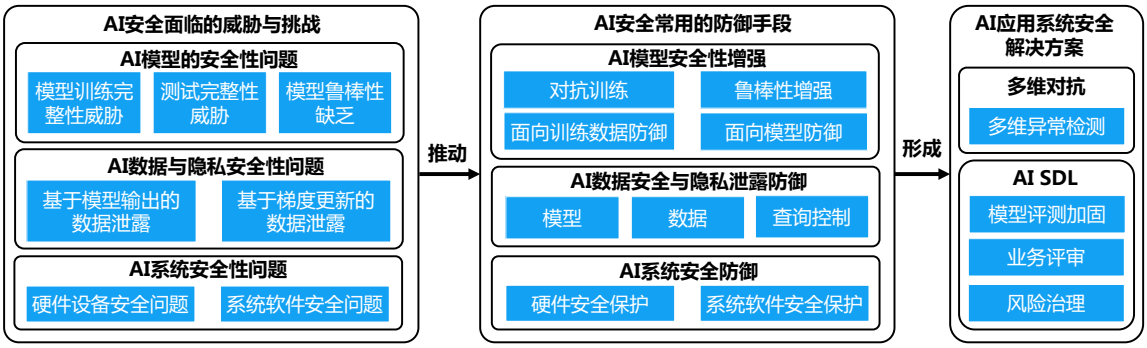


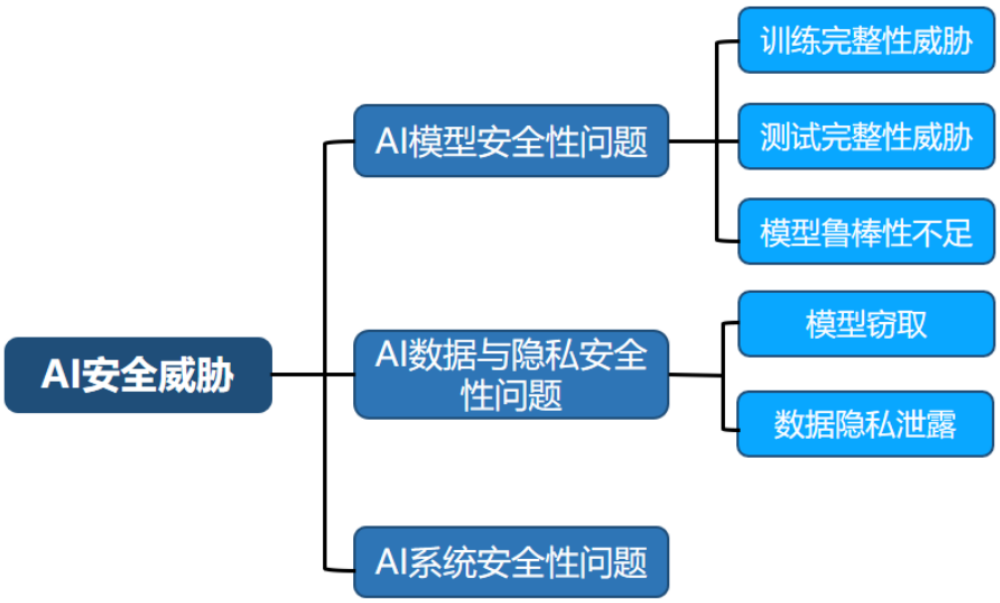
图 2.1: AI 技术面临的安全威胁与挑战、AI 安全常用防御技术以及 AI 应用系统安全解决方案之间的关系

以通过 AI 模型构建和使用过程中产生的信息在一定程度上窃取 AI 模型的数据，主要通过两种方式来进行攻击：1）基于模型的输出结果，模型的输出结果隐含着训练/测试数据的相关属性。以脸部表情识别为例，对于每张查询的输入图片，模型会返回一个结果向量，这个结果向量可能包含关于脸部内容的信息，例如微笑、悲伤、惊讶等不同表情的分类概率，而攻击者则可以利用这些返回的结果信息，构建生成模型，进而恢复原始输入数据，窃取用户隐私 [16]；2）基于模型训练产生的梯度，该问题主要存在于模型的分布式训练中，多个模型训练方之间交换的模型参数的梯度也可被用于窃取训练数据。

- **AI 承载系统安全问题** 承载 AI 技术的应用系统主要包括 AI 技术使用的基础物理设备和软件架构，是 AI 模型中数据收集存储、执行算法、上线运行等所有功能的基础。应用系统所面临的安全威胁与传统的计算机安全威胁相似，会导致 AI 技术出现数据泄露、信息篡改、服务拒绝等安全问题。这些问题可以归纳为两个层面：1）软件框架层面，包含主流的 AI 算法模型的工程框架、实现 AI 技术相关算法的开源软件包和第三方库、部署 AI 软件的操作系统，这些软件可能会存在重大的安全漏洞；2）硬件设施层面，包含数据采集设备、GPU 服务器、端侧设备等，某些基础设备缺乏安全防护容易被攻击者侵入和操纵，进而可被利用施展恶意行为。

图2.1详细描述了 AI 技术面临的安全威胁与挑战、AI 安全常用防御技术以及 AI 应用系统安全解决方案之间的关系，例举了 AI 技术在应用过程中存在的安全威胁和防御技术的种类。在接下来的章节中，我们会全面介绍目前 AI 技术所面临的安全挑战，以及在现实场景中可能出现的安全隐患。

第三章　AI 技术面临的三大威胁域



3.1　AI 模型安全性问题

3.1.1　模型训练完整性威胁

AI 模型的决策与判断能力来源于对海量数据的训练和学习过程。因此，数据是模型训练过程中一个非常重要的部分，模型训练数据的全面性、无偏性、纯净性很大程度上影响了模型判断的准确率。一般来说，一个全面的、无偏的、纯净的大规模训练数据可以使模型很好地拟合数据集中的信息，学习到近似于人类甚至超越人类的决策与判断能力。例如：ImageNet 数据集使 AI 模型在图像分类任务中取得的准确率超越了人类感官判断。但是，如果训练数据受到攻击者的恶意篡改，那么模型将学习到错误的预测能力。例如：在分类模型中，攻击者通过篡改训练数据集中特定样本的标签，导致模型测试阶段针对这些样本输出攻击者指定的标签。这类由数据全面性、无偏性、纯净性引起的安全威胁本质上破坏了模型的训练过程，使模型无法学习到完整的决策、判别能力。因此，在白皮书中，我们也将这类由数据引起的威胁归为破坏模型训练完整性的威胁。破坏模型训练完整性的攻击主要为数据投毒

表 3.1: 攻击方法概括

威胁类型	攻击种类	攻击方法	代表性论文	本文介绍章节
模型训练完整性威胁	数据投毒攻击	标签翻转攻击	[1, 17]	3.1.1
		标签不变攻击	[18, 19]	3.1.1
	后门攻击	常规后门攻击	[2, 20]	3.1.1
		隐蔽后门攻击	[21, 22]	3.1.1
模型测试完整性威胁	对抗攻击	基于扰动的对抗攻击	[3] [23] [24] [25] [26] [27] [28]	3.1.2
		非限制性对抗攻击	[29] [30] [31] [14]	3.1.2
	伪造攻击	面部识别伪造攻击	[32] [33]	3.1.2
AI 数据与隐私安全性问题	基于模型输出的数据泄露	模型窃取攻击	[66]	3.2.1
		模型逆向攻击	[7] [34] [8] [35]	3.2.1
		成员推断攻击	[10],[71],[72],[73],[74],[75]	3.2.1
	基于梯度更新的数据泄露	隐私推断	[36]	3.2.2
		数据生成	[37, 38][39]	3.2.2

攻击 [1], 根据投毒的方法与类型，投毒攻击又可以进一步分为目标固定攻击与后门攻击。接下来，我们将简单介绍投毒攻击、目标固定投毒攻击与后门攻击。

数据投毒攻击

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或协助攻击者在没有破坏模型准确率的情况下入侵模型。前者破坏模型的可用性，为无目标攻击；后者破坏模型的完整性，为有目标攻击。数据投毒攻击最早由 Dalvi 等人在文献 [1] 中提出，他们利用该攻击来逃避垃圾邮件分类器的检测。后来，相关研究人员相继在贝叶斯分类器 [40] 和支持向量机 [41] 等机器学习模型中实现了数据投毒攻击。

破坏完整性的投毒攻击具有很强的隐蔽性：被投毒的模型对干净数据表现出正常的预测能力，只对攻击者选择的目标数据输出错误结果。这种使 AI 模型在特定数据上输出指定错误结果的攻击会导致巨大的危害，在某些关键的场景中会造成严重的安全事故。因此，我们在白皮书中对投毒攻击进行了深入的分析探索，希望这部分内容对读者有所启发。根据攻击者在对毒化模型进行测试时是否修改目标数据，可以将这类攻击分为：目标固定攻击和后门攻击。

目标固定攻击

目标固定攻击是投毒攻击的一种。在这类攻击中，攻击者在模型的正常训练集 $\mathcal{D}_c = (X_c, Y_c)$ 中加入精心构造的毒化数据 $\mathcal{D}_p = (X_p, Y_p)$，使得毒化后的模型将攻击者选定的数据 x_s 分类到目标类别 y_t，而不影响模型在正常测试集的准确率。构造毒化数据 $\mathcal{D}_p$ 的过程可以看作是一个双层优化（Bi-level Optimizaion）的问题。其中，外层优化得到毒化数据 X_p^* 表示如下：

$$X_p^* = \arg\min_{X_p} \mathcal{L}_{\text{adv}}\left(x_t, y_{\text{adv}}; \theta^*\right) \tag{3.1}$$

其中 $\mathcal{L}_{adv}$ 表示攻击者攻击成功的损失，θ^* 表示在 $X_c \cup X_p$ 上训练得到的毒化模型，内层优化得到毒化模型 θ^* 表示如下：

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{\text{train}}\left(X_c \cup X_p, Y; \theta\right) \tag{3.2}$$

可以看到目标梯度 $\nabla_{X_p}\mathcal{L}_{adv}$ 同时由内外层损失函数决定。由于 AI 模型的目标函数是非凸化函数，上述的双层优化问题无法直接求解。

Muñoz-González 等人在 [17] 中首先实现了针对深度神经网络的数据投毒攻击，他们使用反向传播中的梯度优化（Back-gradient Optimization）技术来快速且高效地求解上述的双层优化问题。具体而言，他们通过对内层进行 T 轮迭代展开并优化得到 θ^*，其中每一轮的反向传播都会计算并更新外层优化所需要的梯度 dX_p。然后利用内层优化得到的 θ^* 来计算 $\nabla_{X_p}\mathcal{L}_{adv}$，并与 dX_p 求和得到最终的目标梯度，从而优化得到的 X_p^*，并将 Y_p 的标签翻转到目标类别 Y_t。

上述基于标签翻转的数据投毒攻击可以显式地改变模型的决策边界，这种方法虽然简单且高效，但会导致数据与类别标签不对应。模型训练者会把这种不对应的数据当作异常点从数据集中剔除。因此，Shafahi 等人在 [18] 中提出了标签不变（Clean-Label）的数据投毒攻击，该方法去除之前工作中攻击者可以控制训练数据标签的假设，使得攻击假设更加符合实际场景。在文献 [18] 中，Shafahi 等人采取特征碰撞（Feature Collisions）的方法来优化毒化数据。具体来说，他们通过优化在特征空间上与目标类别图片一致的毒化数据，与此同时保证毒化数据在输入空间上与毒化前尽可能地相似，如下式所示：

$$p = \arg\min_{\mathbf{x}} \|f(x) - f(t)\|_2^2 + \beta\|x - b\|_2^2 \tag{3.3}$$

其中，$f(\cdot)$ 代表模型在倒数第二层（在 Softmax 层之前）的特征空间，t 代表攻击者的目标数据，b 代表优化前的毒化数据。上述的攻击能够隐蔽地影响模型的决策边界，使得攻击者可以在较小比例的毒化数据下完成攻击。

文献 [17] 和 [18] 都基于攻击者完全了解被攻击模型的白盒场景，所以该攻击方法只对特定场景有效。当受害者使用相同的数据替换模型重新训练后，攻击效果便

会失效。因此 Zhu 等人在 [19] 中实现了黑盒场景中的标签不变攻击，即：攻击者不了解受害模型而只能获取到与受害者相似的训练数据集。攻击者使用相似训练数据得到替代模型，并在此基础上优化毒化数据，使得毒化数据具有迁移性，即能够从替代模型传递到受害模型从而实现攻击。具体而言，他们将 [18] 中提出的特征碰撞攻击拓展到凸组合下的集成模型 $\{\sum_{j=1}^{k} c_j^{(i)}\phi^{(i)}\}_i^m$ 下，其中 m 为模型个数，凸组合系数 $c_1, ..., c_k \geq 0$ $\sum_{j=1}^{k} c_j^{(i)} = 1$，具体的优化目标如下：

$$\begin{aligned} \min_{\left\{\mathbf{c}^{(i)}\right\},\left\{\boldsymbol{x}_p^{(j)}\right\}} & \frac{1}{2}\sum_{i=1}^{m} \frac{\left\|\phi^{(i)}(\boldsymbol{x}_t) - \sum_{j=1}^{k} c_j^{(i)}\phi^{(i)}\left(\boldsymbol{x}_p^{(j)}\right)\right\|^2}{\left\|\phi^{(i)}(\boldsymbol{x}_t)\right\|^2} \\ \text{s.t} & \sum_{j=1}^{k} c_j^{(i)} = 1, c_j^{(i)} \geq 0, \forall i, j \\ & \left\|\boldsymbol{x}_p^{(j)} - \boldsymbol{x}_b^{(j)}\right\|_\infty \leq \epsilon, \forall j \end{aligned} \tag{3.4}$$

其中，$\boldsymbol{x}_p^{(j)}$ 代表第 j 个毒化数据，$\boldsymbol{x}_b^{(j)}$ 则为第 j 个优化前的数据。目标函数表示归一化后的毒化数据与目标数据的特征距离，使得目标数据尽可能地接近毒化数据在 m 个集成模型的特征空间的凸多边形周围，由于系数 $\{c_j^{(i)}\}$ 并不固定，这保证了毒化数据在不同模型上的迁移性。

上述的数据投毒攻击考虑的都是外包训练以及迁移学习的场景，攻击者可以直接修改被攻击模型的训练数据。在联邦学习 [42] 等多方计算的应用场景中，为了保护用户的隐私数据，用户的数据只对自己可见，其与服务器之间的交互主要通过模型增量（梯度）进行传输。正是由于这种特殊性，联邦学习中的数据投毒攻击也被称作模型投毒（Model Poisoning）攻击。Bhagoji 等人首先在 [43] 中实现针对联邦学习场景的模型投毒攻击。他们假设联邦学习场景中共有 k 个用户，其中仅存在一个恶意用户 m。恶意用户 m 在第 t 轮尝试向服务器提交毒化模型增量 δ_m^t 来攻击服务器端的全局模型 $\mathbf{w}_G^t$，其优化目标为：

$$\underset{\delta_m^t}{\operatorname{argmin}} \lambda\mathcal{L}\left(\{\mathbf{x}_i, \tau_i\}_{i=1}^{r}, \hat{\mathbf{w}}_G^t\right) + \mathcal{L}\left(\mathcal{D}_m, \mathbf{w}_m^t\right) + \rho\left\|\boldsymbol{\delta}_m^t - \overline{\boldsymbol{\delta}}_{\text{ben}}^{t-1}\right\|_2 \tag{3.5}$$

其中，$\{\mathbf{x}_i, \tau_i\}_{i=1}^{r}$ 为恶意用户的目标数据集，$\hat{\mathbf{w}}_G^t$ 为恶意用户对第 t 轮的全局模型的估计，$\mathcal{D}_m$ 为恶意用户的测试集，$\mathbf{w}_m^t$ 为第 t 轮恶意用户的本地模型，$\overline{\boldsymbol{\delta}}_{\text{ben}}^{t-1}$ 为对其它 $k-1$ 个正常用户在第 $t-1$ 轮提交模型增量的平均值。目标函数的第一项是恶意用户的攻击目标，由于服务器端会将所有用户提交的模型增量进行聚合，这会削弱恶意用户提交的模型增量对全局模型 $\mathbf{w}_G^t$ 的影响，因此使用超参数 λ 用于放大恶意用户提交给服务器的毒化模型增量；目标函数的第二项用于保证恶意用户本地模型在正常数据集上的性能；目标函数的第三项用于保证恶意模型增量与正常模型增量相似，从而能绕开服务器端的检测机制。

后门攻击

在这类攻击中，攻击者在模型的正常训练集 $D_c = (X_c, Y_c)$ 中加入精心构造的毒化数据集 $D_p = (X_p, Y_p)$，使得毒化后的模型将加入攻击者选定的后门触发器（Back-

door Trigger）的数据分类到攻击者的目标类别 y_t，而不影响模型的正常性能。以图像分类为例，攻击者在测试阶段在原图片 x_i 上添加一个具体的图案或扰动作为后门触发器 Δ，具体的过程如下所示：

$$x_i + \Delta = x_i \odot (1 - m) + \Delta \odot m \tag{3.6}$$

其中，$\odot$ 表示元素积，m 代表图像掩码。m 的大小与 x_i 和 Δ 一致，值为 1 表示图像像素由对应位置 Δ 的像素取代，而 0 则表示对应位置的图像像素保持不变。攻击者发动后门攻击的目标可以表示为式3.7：

$$\min \sum_{x \in \mathcal{X}} \ell\left(y_t, f_{\theta^*}(x + \Delta)\right) \tag{3.7}$$

其中，$\mathcal{X}$ 表示模型输入空间的所有数据，θ^* 表示受害者使用毒化后的数据训练得到的模型参数，训练过程的目标函数如式3.8所示：

$$\min_{\theta} \sum_{(x_c, y_c) \in D_c, (x_p, y_p) \in D_p} \ell\left(y_c, f_\theta(x_c)\right) + \ell\left(y_p, f_\theta(x_p)\right) \tag{3.8}$$

其中，f 代表模型结构，θ 代表模型参数，ℓ 代表损失函数。式3.8可以看作是多任务学习（Multi-task Learning）。第一项代表模型在正常任务上的损失函数，这与 D_c 有关；第二项代表攻击者想要模型额外训练的后门任务上的损失函数，而这取决于 D_p。所以后门攻击的关键在于构造合适的 D_p，在经过受害者的训练后门任务后，达到式3.7中的目标。

后门攻击最早由 Gu 等人在 [2] 中提出，他们通过构造含有后门触发器的毒化数据，并将这些数据的标签翻转为 y_t，从而在外包训练和迁移学习中实现该攻击。近些年来，研究人员从数据投毒的方式和攻击的场景等方面对 [2] 中的后门攻击进行改进和拓展，使得攻击变得更加普遍且隐蔽。

改进数据投毒的方式. Liu 等人在 [20] 中实现在白盒场景下对预训练模型的后门攻击。具体而言，他们首先在预训练模型中选择某个中间层，并选择与上一层连接权重较大的 k 个神经元作为后门特征嵌入的位置，然后对后门触发器进行优化使选择的神经元激活值尽可能地大，如式3.9所示：

$$\arg\min_{\Delta} \sum_{i=1}^{k} (tv_i - f_{n_i}(x + \Delta))^2 \tag{3.9}$$

其中，x 表示像素值为 0 的纯黑图像，tv_i 表示第 i 个神经元激活值的目标，f_{n_i} 表示选定的第 i 个神经元的实际值激活值。上述优化得到的后门触发器能够降低毒化数据占正常数据的比例并提高数据毒化攻击的效果。

由于攻击者将图案形式的后门触发器添加在毒化数据中，这很容易被人发现。为了提高后门触发器的隐蔽性，Liao 等人在 [44] 中使用不易被人察觉的扰动作为后门

触发器，他们使用静态扰动和目标自适应扰动生成攻击所需的后门触发器。静态扰动是攻击者预先选择的噪声，而目标自适应扰动则是通过 [45] 中的通用对抗扰动生成的噪声。与传统的数据投毒类似，研究人员希望实现标签不变的后门攻击，从而提高攻击的隐蔽性。Turner 等人在 [46, 47] 中分别基于生成对抗网络以及对抗样本实现了两种标签不变的后门攻击。Saha 等人在 [21] 中基于 [44] 和 [18] 提出一种更加隐蔽的后门攻击：在该场景中，受害者使用攻击者提供的毒化数据对预训练模型进行微调，毒化数据中不含有攻击者选择的后门触发器并且毒化数据的标签没有进行翻转。具体而言，以受害者将预训练模型微调为二分类器（只含有源类别和目标类别）为例，攻击者选择目标类别 y_t 中的 K 张图片 t_k，并使得毒化数据与源类别中添加后门触发器的数据在模型的特征空间尽可能地接近，如此使得毒化数据在输入空间与目标类别尽可能地接近，如式3.10所示：

$$\begin{aligned} &\arg\min_{z} \textstyle\sum_{k=1}^{K} \left\| f\left(z_k\right) - f\left(\tilde{s}_{a(k)}\right) \right\|_2^2 \\ &\text{s.t.} \quad \forall k \quad \left\| z_k - t_k \right\|_\infty < \epsilon \end{aligned} \tag{3.10}$$

其中，$f(\cdot)$ 表示模型的特征空间，$\tilde{s}_{a(k)}$ 表示源类别中第 k 个图像在添加相同图案但不同位置后门触发器后，与 z_k 在模型特征空间上距离最近的一个。公式3.10是一个含有约束的优化问题，可以使用投影梯度下降 (Projected Gradient Descent, PGD) 算法 [24] 进行迭代优化。随后攻击者在测试阶段使用预先选定的后门触发器添加在源类别图像上任意的位置，从而使得毒化模型误分类为目标类别。

拓展后门攻击的场景. 大部分的后门攻击 [2, 20, 21, 44, 46, 47] 都是将数据作为后门的载体，这些攻击需要受害者使用毒化数据训练后才会触发。但是随着预训练模型的广泛使用，将模型作为后门的载体也成为了另外一种攻击的思路。Yao 等人在 [13] 中提出在迁移学习场景下的潜在后门（Latent Backdoor）攻击，攻击者在预训练模型中提前嵌入后门，当受害者在本地对含有攻击者指定目标类别的训练任务进行微调后，后门才会被触发。攻击者首先对目标类别 y_t 生成特定的后门触发器 Δ，其优化目标如公式3.11所示：

$$\arg\min_{\Delta} \sum_{x \in X_{\setminus y_t} \cup X_{y_t}} \sum_{x_t \in X_{y_t}} D\left(f_\theta^{K_t}(x+\Delta)), f_\theta^{K_t}\left(x_t\right)\right) \tag{3.11}$$

其中，X_{y_t} 表示目标类别的数据，$X_{\setminus y_t}$ 表示除目标类别的所有数据，K_t 表示受害者从预训练模型中迁移的最后一层，$f_\theta^{K_t}$ 表示第 K_t 层的特征空间。公式3.11使得任何添加后门触发器的毒化数据在特征空间上尽可能地和正常的目标样本相似。随后，攻击者在正常训练模型的同时使用优化得到的后门触发器进行后门训练，训练的损失函数如公式3.12所示：

$$J_\theta(\theta; x, y) = \ell\left(y, f_\theta(x)\right) + \lambda \cdot D\left(f_\theta^{K_t}\left(x+\Delta\right), \phi_\theta\right) \tag{3.12}$$

其中，$x \in X_{\setminus y_t} \cup X_{y_t}$，其对应标签为 y；$\phi_\theta = \underset{\phi}{\operatorname{argmin}} \sum_{x_t \in X_{y_t}} D\left(\phi, f_\theta^{K_t}\left(x_t\right)\right)$ 表示目标数据在第 K_t 层上的最小特征值。公式3.12的第一项为正常训练的损失函数，第二

项表示毒化数据与目标类别数据在第 K_t 层特征空间的距离，λ 为超参数。当受害者使用含有后门的预训练模型进行微调后，若受害者的目标模型含有 y_t 这一类，攻击者之前嵌入的后门便会触发。

与传统数据投毒攻击类似，联邦学习 [42] 等多方合作训练的场景也容易受到恶意用户发起的后门攻击。Bagdasaryan 等人首先在 [48] 中实现联邦学习场景下的后门攻击，他们的方法与 [43] 中的模型投毒类似，都需要对恶意用户提交的模型增量进行放大来实现对全局模型的攻击。与之不同的是，他们选择在联邦学习训练快收敛的阶段进行攻击，从而可以降低恶意用户攻击的次数并增强攻击的持久性。Xie 等人在 [49] 中提出了联邦学习中的分布式后门攻击。他们假设联邦学习的用户中存在 M 个恶意用户，全局后门触发器被切分为 M 个部分作为 M 个恶意用户的本地后门触发器，每个恶意用户使用本地后门触发器以及共同的目标类别使用 [48] 的方法对全局模型进行后门攻击。这种攻击不仅能够使得攻击更持久而且能保证攻击的隐蔽性。

除了针对图像分类任务的后门攻击，研究人员将后门攻击拓展到 LSTM 等序列模型 [50]、语言预训练模型 [51] 和视频识别模型 [22]。

3.1.2 测试完整性威胁

模型测试阶段是指模型训练完成之后，模型参数被全部固定，模型输入测试样本并输出预测结果的过程。在没有任何干扰的情况下，AI 模型的准确率超乎人们的想象，在 ImageNet 图像分类任务中，识别准确率已经超过了人类。但是，近些年来的研究表明：在模型测试阶段，AI 模型容易受到测试样本的欺骗从而输出不可预计的结果，甚至被攻击者操纵。我们将这类威胁 AI 模型测试阶段正确性的问题定义为测试完整性威胁。**对抗攻击**与**伪造攻击** (Adversarial Attack or Evasion Attack) 是破坏模型测试完整性的典型威胁，本章重点关注对抗攻击与伪造攻击。

对抗攻击

对抗攻击是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。误判既包括单纯造成模型决策出现错误的无目标攻击，也包括受到攻击者操纵导致定向决策的有目标攻击。对抗攻击最早由 Szegedy 等人提出，他们在最基本的图像分类任务中，向分类图像的像素中加入微小的扰动，使得分类模型的准确率严重下降，同时对抗样本具有很强的隐蔽性，攻击者做出的修改往往并不会引起人们的察觉。

这类威胁来自于 AI 模型算法本身的缺陷，广泛存在于 AI 技术应用的各个领域之中，一旦被攻击者利用会造成严重的安全危害。例如：在自动驾驶中，对交通标

志的误识别会造成无人汽车做出错误决策引发安全事故。对抗样本的发现严重阻碍着 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，近些年来对抗攻击以及其防御技术吸引了越来越多的目光，成为了研究的一大热点，涌现出大量的学术研究成果。接下来，我们会在白皮书中介绍对抗攻击基本原理与攻击设定，主流的攻击技术以及在不同领域中的应用。

对抗攻击原理与威胁模型对抗攻击的基本原理就是对正常的样本添加一定的扰动从而使得模型出现误判。以最基本的图像分类任务为例，攻击者拥有若干数据 $\{x_i, y_i\}_{i=1}^N$，其中 x_i 代表数据集中的一个样本也就是一张图像，y_i 则是其对应的正确类别，N 为数据集的样本数量。将用于分类的目标模型表示为 $f(.)$，则 $f(x)$ 表示样本 x 输入模型得到的分类结果。攻击者应用对抗攻击的方法对正常样本 x 进行修改得到对应的对抗样本 x'，该对抗样本可以造成模型出现误判，同时其与原样本的应该较为接近具有同样的语义信息，一般性定义如下：

$$x' : \|x - x'\|_D < \epsilon, f(x') \neq y \tag{3.13}$$

其中 $\|.\|_D$ 代表着对抗样本与原样本之间的某种距离度量，为了使修改的样本能够保持语义信息不造成人类的察觉，两者之间的距离应该足够小，同时造成最后模型判断出现错误，分类结果不同于正确类别，而 ϵ 就是对抗样本与原样本之间设定的最大距离，其取值往往和具体的应用场景有关。

根据攻击意图，对抗攻击可以分为有目标攻击和无目标攻击。以上的一般定义属于无目标攻击，即经过修改的样本只要造成错误使得分类标签与原标签不同即可；有目标攻击是指攻击者根据需要对样本进行修改，使得模型的分类结果变为指定的类别 t，定义如下：

$$x' : \|x - x'\|_D < \epsilon, f(x') = t \tag{3.14}$$

根据攻击者所能获取的信息，对抗攻击可以分为黑盒攻击和白盒攻击。黑盒攻击是指攻击者在不知道目标模型的结构或者参数的情况下进行攻击，但是攻击者可以向模型查询特定的输入并获取预测结果；白盒攻击是指攻击者可以获取目标模型 $f_\theta(.)$ 的全部信息，其中 θ 代表模型的具体参数，用于实施有针对性的攻击算法。一般情况下，由于白盒攻击能够获取更多与模型有关的信息，其攻击性能要明显强于对应的黑盒攻击。以上我们对攻击的主要目标与攻击设置进行了简要的介绍，在不同设置下各种攻击具有不同的特点，主流的攻击技术可以分为基于扰动的对抗攻击和非限制对抗攻击。

基于扰动的对抗攻击　最初的对抗攻击算法主要是基于扰动的对抗攻击，这类攻击在图像分类任务上被广泛研究，也是最主要的攻击类型。这类攻击的主要思想就是在输入样本中加入微小的扰动，从而导致 AI 模型输出误判。以图像分类任务为例，攻击者可以对输入图像的像素添加轻微扰动，使对抗样本在人类看来是一幅带有噪

声的图像。考虑到攻击的隐蔽性，攻击者会对这些扰动的大小进行限制从而避免人类的察觉。已有的研究通常基于扰动的范数大小 ℓ_p 度量样本之间距离

$$\ell_p : \|\delta\|_p = \|x - x'\|_p = (\sum_{k}^{n} |(x_i - x_i')|^p)^{1/p} \tag{3.15}$$

其中 x_i、$x_i^{'}$ 分别指正常样本和对抗样本在第 i 处的特征，在图像任务中为对应位置的像素值。目前对抗攻击算法的主要思想是将生成对抗样本的过程看做一个优化问题的求解。接下来我们首先介绍几种白盒对抗攻击算法，之后介绍一些针对防御技术的攻击增强算法，最后给出几种针对黑盒模型的攻击方法。

白盒攻击算法. 在模型训练时，研究者通过优化模型的参数使得模型的损失函数 $L(f(x), y)$ 最小化；而攻击的过程恰好相反，攻击者希望在模型参数固定的情况下，通过优化输入样本的扰动，使扰动后的数据对模型的损失函数最大化，从而达到错误分类的目的。以应用广泛的 Fast Gradient Sign Method (FGSM) 算法 [23] 为例：

$$x' = x + \epsilon \cdot sign(\nabla_x J(x, y)) \tag{3.16}$$

FGSM 算法沿着目标样本产生梯度符号方向进行单步优化，使扰动朝着增大损失函数最快的方向优化更新。虽然单步优化高效，却容易错过扰动范围内的最优解，因为损失函数在优化空间上并不是一个线性函数。在近年来的研究中，研究者发现多步优化算法 Basic Iterative Method/Projected Gradient Descent (BIM/PGD)[24] 能够进行更好地找到全局最优点，其优化目标如下：

$$x'_t = \pi(x'_{t-1} + \alpha \cdot sign(\nabla_x J(x'_{t-1}, y))) \tag{3.17}$$

其中，x'_t 是指经过 t 轮优化之后得到的样本，α 是指多步优化中每一次优化的步长。PGD 等多步优化算法会对样本进行多轮优化，一旦扰动超过预先设定的最大值，则进行投影操作 π 将扰动噪声约束到有效的限制范围内。除了在设定好的扰动范围内寻找损失函数的极大值点，还可以在增大分类损失函数的同时尽可能地减小对抗扰动本身的大小，以生成有效而隐蔽的对抗样本。C&W 算法 [25] 重新设计了攻击的损失函数，在增大目标模型分类误差的同时减小对抗样本与原样本之间的距离，定义如下式所示：

$$\min_{\delta} \|\delta\|_p + \lambda \cdot J_{C\&W}(x + \delta, y), \quad s.t. x + \delta \in [0, 1] \tag{3.18}$$

损失函数的第一项用于最小化对抗扰动的大小，$J_{C\&W}(.)$ 是该攻击算法中自定义的损失函数，用于衡量分类的误差。我们可以发现无论上述算法的具体形式如何变化，都遵循对抗攻击的基本原理：通过优化算法对图片像素直接添加微小的扰动来最大化目标模型的损失函数，从而使经过修改的样本结果偏离真实值，达到攻击的效果。其它不同形式的基于白盒优化的攻击算法包括 Distributionally Adversarial Attack[26], Jacobian-based Saliency Map Approach[27]，Elastic-net Attack[28] 等。

针对防御的攻击增强. 白盒攻击场景中，模型很难抵御对抗攻击，即便使用了某些对抗防御技术，模型依然会被适应性的增强攻击算法所攻破。Athalye 等人在研究中 [52] 提到，基于白盒优化的攻击算法在施展的时候需要有效的梯度，因此随后提出的防御技术大多基于混淆梯度的思想，例如梯度破碎，在网络中加入不可微的操作从而阻止攻击者获取有效的梯度，可以表示为 $\hat{f}(x) = f(g(x))$，其中 $g(x)$ 是某种不可微不平滑的预处理，因此无法通过反向传播得到有效的梯度 $\nabla_x \hat{f}(x)$；随机梯度，在输入样本进入网络前进行随机变换或者对网络添加某些随机操作从而干扰真实梯度的计算，可以表示为 $\hat{f}(x) = f(t(x))$，其中 $t(x)$ 表示对具有一定分布 T 的随机变换，求得的梯度 $\nabla_{t(x)} \hat{f}(t(x))$ 经过扰动有效性就会降低；梯度爆炸或消失，在模型预测时加入某些复杂操作 $g(x)$，这些操作往往涉及多步优化过程，造成对 $f(g(x))$ 求导过程出现梯度爆炸或者梯度消失的现象，导致优化无法正常进行。

然而基于上述对抗防御技术被证明在针对性的白盒增强攻击下会被轻易攻破 [52]：针对梯度破碎，可以使用 Backward Pass Differential Approximation(BPDA)，通过一个可微的函数来模拟不可微的部分，从而获取有效的梯度，例如函数不可微部分 g 往往具有如下性质 $g(x) \approx x$，因此我们就可以模拟 $f(g(x))$ 在样本 $\hat{x}$ 处的偏导 $\nabla_x f(g(x))|_{x=\hat{x}} \approx \nabla_x f(x)|_{x=g(\hat{x})}$，攻击过程中只需要正常进行向前传播，而在反向传播的过程中，我们将不可微层 $g(x)$ 替换为恒等映射，利用上式模拟真实的梯度进而完成优化；针对随机梯度，Expectation over Transformation(EOT) 攻击算法被提出，用于解决在一定分布 T 的随机变换 $t(.)$ 下的优化问题，即优化在一定随机变换下的期望 $E_{t\sim T} f(t(t))$ 并使用梯度下降法求解，可以根据 $\nabla E_{t\sim T} f(t(x)) = E_{t\sim T} \nabla f(t(x))$ 求得每一步梯度下降时样本在一定随机变换下得到的梯度期望，进而有效地优化该问题，除了应对随机化的防御机制，EOT 攻击还可以用于应对物理场景下的各种随机因素 [53]；针对梯度消失和复杂化，Athalye 等人提出重参数化的攻击方法，针对复杂的优化过程 $g(x)$ 寻找一类重参数化的映射 $x = h(z)$ 使得 $g(h(z)) = h(z)$，例如 $g(x)$ 可以表示使用对抗生成网络的生成器生成最接近 x 的样本 x_{gan}，因而样本 x 可以重新使用潜空间的变量 z 重参数化表示为 $x = h(z)$，那么 $h(z)$ 本身就是生成器可以生成的样本即 $g(h(z)) = h(z)$，因而对原样本进行的优化可以转化为在潜空间上的优化，而 $\nabla_z f(h(z))$ 在白盒模式下是可以有效计算得到的，进而可以有效地生成对抗样本 $h(z)$。

黑盒攻击算法. 通过上述的介绍我们可以得出结论：在完全白盒的情况下，攻击者可以根据防御的特点进行适应性的攻击调整，提高攻击效果。但在黑盒模式下攻击者缺少模型的结构参数等信息，同时也不了解模型采取的防御手段，其发起的对抗攻击的有效性会被大大削弱。模型信息的缺失导致攻击者无法获取模型损失在对应样本上的梯度信息，进而使其难以进行有效的优化与对抗样本搜索。黑盒攻击面临的主要问题就是如何在这种情况下对正常样本进行优化。经过研究者的一些发现和尝试，黑盒攻击主要可以从两个角度来施展：1) 基于迁移性的攻击，对抗样本

的传递性是指针对某种模型在白盒模式下生成的对抗样本，输入其它黑盒模型之后同样具有一定的攻击效果，这是由于不同模型在相同的任务中会学习到相似的特征，基于这种性质攻击者可以先自行训练一个具有相同功能的替代模型，针对这个白盒的替代模型生成对抗样本用以攻击黑盒的目标模型，此时可以选用上面介绍的白盒算法。攻击者可以通过一些手段增强对抗样本的传递性，例如：Papernot 等人提出使用模型窃取的方法，增强替代模型与目标模型决策边界的相似性，从而加强对抗样本针对目标模型的攻击性 [54]；Dong 等人提出在多步优化算法的基础上，引入优化中动量的概念，使得每一步优化的方向不单单由当前的梯度方向决定，而是和历史积累的前进方向共同决定，减少样本陷入某个局部极值的可能性，从而增强对抗样本在模型之间的传递性，基于该方法的攻击获得了 NIPS 2017 黑盒对抗攻防竞赛的第一名 [55]，公式如下：

$$\begin{aligned} g_t &= \mu \cdot g_{t-1} + \frac{\nabla_x J(f(x'_{t-1}), y)}{\|\nabla_x J(f(x'_{t-1}), y)\|_1} \\ x'_t &= x'_{t-1} + \alpha \cdot sign(g_t) \end{aligned} \tag{3.19}$$

其中 μ 是历史优化方向的系数，$\alpha = \frac{\epsilon}{T}$ 是每步优化的步长，T 是多步优化的总步数。2) 基于查询的攻击，这类方法与利用对抗样本的传递性的攻击方式不同，攻击者会向目标模型查询一定的样本，根据目标模型返回的结果近似求解损失函数对当前样本的梯度，以此进行有效的攻击，例如：Chen 等人将零阶优化（Zeroth Order Attack）的方法引入对抗攻击，该方法通过向目标模型查询成对的扰动样本，近似求解损失函数在样本某一维度上的偏导，进而可以近似估算损失函数在该样本上的梯度 [56]，近似计算的过程大致如下：

$$\nabla_{x_i} J(f(x), y) = \frac{J(f(x + h \cdot e_i), y) - J(f(x - h \cdot e_i), y)}{h} \tag{3.20}$$

其中 e_i 是维度 i 方向上的单位向量，h 是在该方向上扰动的步长。

非限制对抗攻击　在早期的对抗样本研究中，为了使得修改过的对抗样本不引起人类的注意，避免样本语义信息的变化，要求只能向样本添加微小的扰动来生成对抗样本。而本小节介绍的非限制对抗攻击是指遵循对抗样本基本定义和原理，但不受扰动条件限制的其它对抗攻击模式。

基于扰动的攻击主要有两点局限性：1）微小的对抗扰动在现实情况中难以实施。例如：在现实图像分类任务中，攻击者只能对目标本身进行修改，无法对背景部分添加相应的对抗扰动，同时这样微小的对抗扰动在环境因素的影响下很容易失效；2）在现实场景中，基于扰动的对抗攻击只是对抗攻击中的一种技术。对抗样本的原始定义为：经过攻击者恶意设计和篡改，在不引起人类察觉的情况下欺骗 AI 模型的样本。而基于扰动的对抗攻击只是这类恶意篡改中的一类，现实的攻击场景则存在着大量潜在的攻击手段。因此可以认为，非限制对抗攻击回归了对抗样本的本质定义，

不再局限于原样本的扰动域，只要保证样本语义信息合理并且能够使模型产生误判即可。

近些年来，研究者提出了多种非扰动限制的攻击方法。例如：利用对抗生成网络生成对抗样本 [29]。对抗生成网络是一类强大的生成模型，学习样本数据分布，接收潜空间输入并生成与训练数据相似的样本。对抗生成网络可以使用 $x_c = h(z|c)$ 来表示，其中 x_c 表示某一类别 c 的生成样本。在基于扰动的对抗攻击中，攻击者对正常样本进行扰动；而在对抗生成网络中，攻击者优化判别损失函数在潜空间 z 上的优化问题，即

$$\max_{z} J(f(h(z|c)), c) \tag{3.21}$$

优化 z 使生成器生成的样本看起来像没有经过扰动的正确类别样本，但实际上目标模型在该样本的判断上误差较大。

添加对抗补丁 (Adversarial Patch) [30] 也是一种常见的现实域攻击手段，这种方法并非在图片所有的位置上添加扰动，而是在图片的某个区域设计特殊的统一对抗图案，当该图案被插入正常图像中时，AI 模型就会受到干扰乃至产生误判，这种方法相比于微小的扰动更容易在现实场景下复现使用。

根据以上所阐述的生成对抗样本的基本方法原理，对抗攻击很快在目标检测 [4]、图像分割 [57]，3D 识别 [58, 59]，语音识别 [60, 61]，强化学习 [62, 63, 64]，自然语言处理 [65, 66] 等多种 AI 任务中获得了成功，这些研究证明不同领域中的 AI 模型普遍受到对抗样本的威胁。在这些研究中，部分攻击已被证明在实际场景下可以生效。例如：在目标检测应用中，攻击者在限速标志或者人身上贴上对抗图案，可以使检测结果发生错误 [53, 4]；在语音识别系统中，攻击者可以模仿受害人的样本以获得相应权限 [67]；在自然语言处理任务中，攻击者通过对文本进行少量修改而欺骗文本检测系统 [66]。

伪造攻击

伪造攻击是向生物识别系统提交伪造信息以通过身份验证的一种攻击方式 *[32]*，是一种 *AI* 测试完整性威胁。生物验证技术包括指纹核身、面容核身、声纹核身、眼纹核身、掌纹核身等等。以声纹核身为例，攻击者有很多种方法来进行伪造攻击声纹识别系统、声纹支付系统、声纹解锁系统等。例如：攻击者对声纹解锁系统播放一段事先录制或者人工合成的解锁音频通过验证。在这类音频伪造攻击中，攻击者可以通过手机等数码设备直接录制目标人物的解锁音频，也可以通过社交网络检索目标账号获取解锁音频。甚至，攻击者可以从目标人物的多个音频中裁剪合成解锁音频，或者通过深度语音合成技术来合成目标人物的解锁音频 [68, 69, 70]。

3.1.3 模型鲁棒性缺乏

实际上，以上两种安全威胁暴露出了 AI 模型自身存在的安全隐患。模型鲁棒性要求 AI 模型对于异常和存在微小扰动的输入样本，能够保持输出稳定、准确的预测结果。鲁棒性缺乏主要原因有两类：1）真实环境因素多变，AI 模型在真实使用的过程中表现不够稳定；2）AI 模型可解释性不足，目前 AI 技术中广泛使用的深度学习模型是由多层神经网络模块连接组合而成，模型参数数量巨大、体系结构复杂，是一个结构复杂、难以使用清晰解析式来表达的非凸函数。因此，即便是在没有遭遇恶意攻击的情况下，也可能出现预期之外的安全隐患。

环境因素多变 在真实场景下，AI 模型往往存在鲁棒性不足的问题，投入使用后的准确率不及训练、测试时候良好，甚至会出现一些预料之外的错误结果。这种现象的主要原因是训练数据不够充足，AI 模型难以学习到真实场景中的全部情况。在真实场景下，正常的环境因素变化也会对模型的可靠性产生影响。例如：光照强度、视角角度距离、图像仿射变换、图像分辨率等环境因素会对模型产生不可预测的影响。来自 MIT 的研究人员们构建了一个新的验证数据集集 ObjectNet[71]，该数据集包含了 313 类不同的日常物体图片。这些图片是工作人以严格标准精心构建的，图片的变化因素包含：物体摆放的方向、拍摄的角度以及物体放置的背景，例如：物体不能总是正面朝向镜头、杂乱的背景包含复杂的语义信息，这些考虑是为了充分还原现实世界的复杂性。试验中，使用该数据集中与 ImageNet 重合的 113 类物体图片进行测试，模型采用在 ImageNet 上预训练好的主流识别模型，经过试验这些模型在 ObjectNet 数据集上的识别性能明显下降，准确率下降 40%-45%。上述实验说明，即使对于成熟的图像识别任务，AI 技术在面对真实复杂的场景时仍然存在鲁棒性不足的问题。

可解释性缺乏 可解释性是指，在得到模型输出结果的基础上，解释 AI 模型所做决策背后的逻辑以及使人相信其决策准确性的能力。换句话说，就是回答一个“为什么”的问题，即可以解释为什么模型可以通过输入的信息来进行相应的决策。以及在构建 AI 模型的过程中，为什么当前的模型设计可以获得良好的性能。

模型可解释性的发展可以帮助我们更好的理解模型本身，了解模型输入数据是如何影响输出结果，这对于揭示攻击的原理和增强模型的安全性有着重要帮助。在 2018 年 5 月，欧盟的《通用数据保护规范（GDPR)》就要求 AI 算法具有一定的可解释性 [72]。然而 AI 模型是一种由大量的基本操作组合而成的复杂运算，研究者难以对这种复杂结构的行为进行逐一分析，导致 AI 模型成为一个严重缺乏可解释性的黑盒运算。因此，模型参数的微小改变都可能会引起不可预期的预测结果，人类也无法直接理解神经网络是如何操作的。例如：在图像分类任务中模型可以进行非常精确的分类但是却仅仅给出分类的概率，人们不得而知分类结果是如何计算得来的，

无法直接理解模型的行为。

可解释性的缺乏在实际使用中可能会引起很多负面影响：1）模型行为难以预测。由于研究者无法直接理解模型的决策机理，因此当模型在面对复杂多样的现实场景时，研究者就难以对模型的一些意料之外的行为进行预估，进而导致严重的安全隐患。这在自动驾驶等领域会产生难以挽回的后果；2）人们对 AI 技术信任感的缺失。由于无法理解 AI 算法的决策逻辑，人们很难对不透明的 AI 算法产生认同感，这严重阻碍了 AI 技术在金融、医疗、交通等攸关人身财产安全，对安全性、可靠性要求较高的领域中的发展；3）AI 算法设计时缺乏理论根据。如果不理解模型的决策机理以及模型各种架构对性能的影响，构建者在设计相关 AI 算法时就会陷入盲目而混乱的尝试之中，最终得到的算法很有可能来源于有限的性能测试，其性能是如何得到的却无法解释。这不仅仅限制了 AI 模型在多种场景下的泛化能力，还使得模型进行调整、功能迁移、安全加固等操作缺乏指导方向。

3.1.4 模型偏见威胁

尽管 AI 模型在许多分类任务中取得了惊人的效果，但是它却可能带有偏见或歧视。自 2017 年底以来，AI 模型偏见（AI Bias）就成为学术界及人工智能行业疾呼需要解决的问题。目前学术界谈论较多的 AI 模型偏见有性别歧视、种族歧视等。除了这类偏见，在金融领域也存在 AI 模型偏见，例如：银行信用评分偏见、保费计算偏见、保费赔偿决策偏见等。这些偏见会让 AI 模型做出不适宜的判断结果，对行业和社会造成不小的损失。

最近有很多关于模型偏见算法的研究，模型偏见来源于有意识或无意识、文化差异、个人因素、人口均等/不同影响以及机会均衡等。通常来说，AI 模型偏见来源于训练数据的偏差，这些偏差可以归类为以下三种:

心理偏见

心理偏见（Psychophysical Bias）来源于对刺激所产生的决策偏见，一个典型的例子是“夏彭特错觉”（Charpentier Illusion）。在“夏彭特错觉”现象中，针对质量相同的两个物体，人们倾向于低估体积较小的物体的质量。因此，他们会对重量空间中的小尺寸产生偏见。目前，大多数商业 AI 系统使用有监督机器学习，标签用来训练模型、计算模型梯度更新。通常情况下，数据收集人员人工对训练数据打上标签。然而由于人们经常表现出心理偏见，这些心理偏见会影响到训练数据的客观事实。如果 AI 模型被训练出来用于估计这些标签，这种对特定对象不公平的分类将被编码到 AI 模型，最终将导致 AI 模型产生偏见。

歧视性偏见

歧视性偏见（Discriminatory Bias）主要包含对种族、性别等歧视。由于收集数据时候存在歧视性偏见，这将导致使用这些数据训练的 AI 模型可能会继承歧视性偏见。政府和公司在使用包含歧视性偏见的 AI 模型时，实际上对于被检测对象是不公平的。但是，关于 AI 模型的偏见是否违反反歧视法律，则需要进行人工判断或进行外部验证。

统计偏见

统计偏见（Statistical Bias）是指训练数据集的分布和实际数据的分布之间存在差异。当训练数据中缺失某些类型的数据，或者训练数据分布不均衡的时候，就会导致训练数据集存在统计偏见。而使用缺失完整性数据集训练出来的 AI 模型，对于存在偏见的测试数据则不能均衡的输出公平的预测结果。在这种情况下，除了增加训练数据的多样性以外，还可以通过修改模型结构、增加数据预处理等方法防止模型过拟合，消除模型中存在的偏见。例如：深度神经网络通常使用数据增强来预处理输入数据；增加 Dropout 层、引入正则函数等方法来消除由统计偏见导致的 AI 模型偏见。

3.2 AI 数据与隐私安全性问题

由于 AI 技术使用过程中产生的模型梯度更新、输出特征向量以及预测结果与输入数据、模型结构息息相关，因此 AI 模型产生的计算信息面临着潜在的隐私数据泄露、模型参数泄露风险。

3.2.1 基于模型输出的数据泄露

在 AI 模型测试阶段，AI 模型参数被固定住，测试数据输入模型并输出特征向量、预测结果等信息。例如：在图像分类任务中，模型的输出包含卷积层输出的特征向量、Softmax 层输出的预测概率向量等。近些年来研究结果表明，模型的输出结果会隐含一定的数据信息。攻击者可以利用模型输出在一定程度上窃取相关数据，主要可以窃取两类数据信息：1）模型自身的参数数据；2）训练/测试数据。

模型窃取

模型窃取攻击 (Model Extraction Attack) 是一类隐私数据窃取攻击，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时间构建而成的，对拥有者来说具有巨大的商业价值。一旦模型的信息遭到泄露，攻击者就能逃避付费或者开辟第三方服务，从而

获取商业利益，使模型拥有者的权益受到损害。如果模型遭到窃取，攻击者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增加攻击的成功率，造成严重的安全风险。

目前，大多数 AI 技术供应商将 AI 应用部署于云端服务器，通过 API 来为客户端提供付费查询服务。客户仅能通过定义好的 API 向模型输入查询样本，并获取模型对样本的预测结果。然而即使攻击者仅能通过 API 接口输入请求数据，获取输出的预测结果，也能在一定情况下通过查询接口来窃取服务端的模型结构和参数。模型窃取攻击主要可以分为三类：1) Equation-solving Attack；2) 基于 Meta-model 的模型窃取；3) 基于替代模型的模型窃取。

Equation-solving Attack 是一类主要针对支持向量机 (SVM) 等传统的机器学习方法的模型窃取攻击。攻击者可以先获取模型的算法、结构等相关信息，然后构建公式方程来根据查询返回结果求解模型参数 [7]。在此基础之上还可以窃取传统算法中的超参数，例如：损失函数中 loss 项和 regularization 项的权重参数 [73]、KNN 中的 K 值等。Equation-solving Attack 需要攻击者了解目标算法的类型、结构、训练数据集等信息，无法应用于复杂的神经网络模型。

基于 Meta-model 模型窃取的主要思想是通过训练一个额外的 Meta Model $\Phi(\cdot)$ 来预测目标模型的指定属性信息。Meta Model 的输入样本是所预测模型在任务数据 x 上的输出结果 $f(x)$，输出的内容 $\Phi(f(x))$ 则是预测目标模型的相关属性，例如网络层数、激活函数类型等。因此为了训练 Meta Model，攻击者需要自行收集与目标模型具有相同功能的多种模型 $f_i(\cdot)$，获取它们在相应数据集上的输出，构建 Meta Model 的训练集。然而构建 Meta Model 的训练集需要多样的任务相关模型，对计算资源的要求过高，因此该类攻击并不是非常实用，而作者也仅在 MNIST 数字识别任务上做了实验 [34]。

基于替代模型训练的是目前比较实用的一类模型窃取攻击。攻击者在未知目标模型结构的情况下向目标模型查询样本，得到目标模型的预测结果，并以这些预测结果对查询数据进行标注构建训练数据集，在本地训练一个与目标模型任务相同的替代模型，当经过大量训练之后，该模型就具有和目标模型相近的性质。一般来说，攻击者会选取 VGG、ResNet 等具有较强的拟合性的深度学习模型作为替代模型结构 [35]。基于替代模型的窃取攻击与 Equation-solving Attack 的区别在于，攻击者对于目标模型的具体结构并不了解，训练替代模型也不是为了获取目标模型的具体参数，而只是利用替代模型去拟合目标模型的功能。为了拟合目标模型的功能，替代模型需要向目标模型查询大量的样本来构建训练数据集，然而攻击者往往缺少充足的相关数据，并且异常的大量查询不仅会增加窃取成本，更有可能会被模型拥有者检测出来。为了解决上述问题，避免过多地向目标模型查询，使训练过程更为高效，研究者提出对查询的数据集进行数据增强，使得这些数据样本能够更好地捕捉目标模型的特点 [8]，例如：利用替代模型生成相应的对抗样本以扩充训练集，研究认为

对抗样本往往会位于模型的决策边界上，这使得替代模型能够更好地模拟目标模型的决策行为 [54, 74]。除了进行数据增强，还有研究表明使用与目标模型任务无关的其它数据构建数据集也可以取得可观的攻击效果，这些工作同时给出了任务相关数据与无关数据的选取组合策略 [75, 35]。

隐私泄露

机器学习模型的预测结果往往包含了模型对于该样本的诸多推理信息。在不同的学习任务中，这些预测结果往往包含了不同的含义。例如：图像分类任务中，模型输出的是一个向量，其中每一个向量分量表示测试样本为该种类的概率。最近的研究结果证明，这些黑盒的输出结果可以用来窃取模型训练数据的信息。例如：Fredrikson 等人提出的模型逆向攻击（Model Inversion Attack）[6] 可以利用黑盒模型输出中的置信度向量等信息将训练集中的数据恢复出来。他们针对常用的面部识别模型，包括 Softmax 回归，多层感知机和自编码器网络实施模型逆向攻击。他们认为模型输出的置信度向量包含了输入数据的信息，也可以作为输入数据恢复攻击的衡量标准。他们将模型逆向攻击问题转变为一个优化问题，优化目标为使得逆向数据的输出向量与目标输出向量差异尽可能地小，也就是说，假如攻击者获得了属于某一类别的输出向量，那么他可以利用梯度下降的方法使得逆向的数据经过目标模型的推断后，仍然能得到同样的输出向量。

成员推断攻击 (Membership-Inference Attack) 是一种更加容易实现的攻击类型，它是指攻击者将试图推断某个待测样本是否存在于目标模型的训练数据集中，从而获得待测样本的成员关系信息。比如攻击者希望知道某个人的数据是否存在于某个公司的医疗诊断模型的训练数据集中，如果存在，那么我们可以推断出该个体的隐私信息。我们将目标模型训练集中的数据称为成员数据 (Member Data)，而不在训练集中的数据称为非成员数据 (Non-member Data)。同时由于攻击者往往不可能掌握目标模型，因此攻击者只能实施黑盒场景下的成员推断攻击。成员推断攻击是近两年来新兴的一个研究课题，这种攻击可以用于医疗诊断、基因测试等应用场景，对用户的隐私数据提出了挑战，同时关于这种攻击技术的深入发展及其相关防御技术的探讨也成为了一个新的研究热点。

2017 年，Shokri[12] 等人第一次提出了成员推断攻击。经过大量实验，他们完成了黑盒场景下成员推断攻击的系统设计。这种攻击的原理是机器学习模型对成员数据 (Member Data) 的预测向量和对非成员数据 (Non-member Data) 的预测向量有较大的差异，如果攻击者能准确地捕捉到这种差异，就可以完成成员推断攻击。然而在黑盒的场景中，我们可以从目标模型中得到的只有预测向量。在实际场景中，由于企业的使用限制，我们也无法从目标模型中获得足够多样本的预测向量。此外，由于不同样本的预测向量的分布本身就不一致，如果攻击者直接利用预测向量进行训练，也无法实现较好的攻击效果。因此，Shokri 等人使用与目标网络同样的结构，并建

立与目标数据集同分布的影子数据集（Shadow Dataset），之后为每一类数据建立多个影子模型（Shadow Model），实现了对预测向量的数据增强效果，并获得了大量的预测向量作为攻击模型的训练样本。最终，Shokri 等人利用预测向量构建了攻击模型，使其能够捕捉预测向量在成员数据和非成员数据之间的差异，从而完成了黑盒场景下的成员推断攻击。

之后随着成员推断攻击技术的发展，人们发现这种攻击的本质就是目标模型对成员数据和非成员数据给出的预测向量存在差异，即成员数据的输出向量的分布更集中，而非成员数据的输出向量的分布相对较为平缓。这种差异性是模型过拟合的主要表现，也就是说成员推断攻击与模型的过拟合程度有很大关联。在这个研究方向上，Samuel Yeom[76] 等人研究了模型的过拟合对成员推理攻击的影响，他们通过理论和实验证实了模型的过拟合程度越强，模型泄露训练集成员关系信息的可能性越大；但同时也指出，模型的过拟合并不是模型易受成员推理攻击的唯一因素，一些过拟合程度不高的模型也容易受到攻击。Ashamed 等人 [77] 进一步完善了黑盒场景下的成员推断攻击，他们在 2019 年提出了改进后的成员推断攻击，在极大地降低了实现这种攻击的成本的同时，实现了与 Shokri 等人 [12] 相同的攻击效果，并更明确地展示了成员推断攻击出现的本质原因，即成员数据和非成员数据的预测向量间的差异主要体现为预测向量的集中度。同时他们提出了三种方法减少了成员推断攻击的部署成本。第一种情况下，他们对目标模型的输出向量从大到小进行重排序，使得模型对不同类别数据的输出向量的分布趋于一致，均为从大到小，这样就可以避免数据增强的过程，进而减少所需 Shadow Model 的数量，同时也不需要知道目标模型的结构，只需要使用基础的网络结构，如 CNN、Logistic Regression 和随机森林等来构建 Shadow Model 即可。同时他们也发现，只需要截取排序后预测向量的前三个概率值并作为新的训练样本，也能达到较好的攻击效果。第二种情况下，他们提出了数据迁移攻击，即使用与目标模型的训练集分布不同的数据集来训练 Shadow Model，最终获得的攻击模型同样能对目标模型的数据进行成员关系推断，并实现类似的攻击效果。第三种情况下，他们提出了 Threshold Choosing，使用该策略可以确定出一个阈值 T，只要预测向量的最大值大于 T，即称该向量对应的待测样本为成员数据，否则判定为非成员数据。Ashamed 等人的工作进一步强化了成员推断攻击，极大地提升了该攻击的威胁性。

随着人们对成员推断攻击研究的深入，研究者们发现了成员推断攻击的一些新特性。如 Shokri 等人 [78] 发现当一个机器学习模型被加入了一些抵御对抗样本攻击的方法后，会提高该模型泄露成员隐私信息的风险。也就是说机器学习模型在对抗样本安全性和成员数据隐私性之间存在一个 trade-off，如果提高了模型抵御对抗样本的能力，同时也会提高从模型中推断出成员数据存在与否的可能，反之亦然。此外，Ahmed Salem 等人 [79] 将成员推断攻击拓展到了在线学习领域。他们发现当机器学习模型完成在线学习后，可以通过更新前后的模型对同一个数据集给出的预测

向量的差异，来完成对目标模型更新集中特定数据的存在性推断，甚至完成对更新集数据的重建。Jamie Hayes[80] 利用生成对抗网络 (GAN) 完成了成员推断攻击的构建。Shokri 等人 [81] 也研究了白盒场景下成员推断攻击，他们利用了成员数据和非成员数据在模型梯度上的差异，完成了推断攻击，并成功绕过了之前提出的一些防御手段，达到了较高的攻击率。

3.2.2 基于梯度更新的数据泄露

梯度更新是指模型对参数进行优化时，模型参数会根据计算产生的梯度来进行更新，也就是训练中不断产生的梯度信息。梯度更新的交换往往只出现在分布式模型训练中，拥有不同私有数据的多方主体每一轮仅使用自己的数据来更新模型，随后对模型参数的更新进行聚合，分布式地完成统一模型的训练，在这个过程中，中心服务器和每个参与主体都不会获得其它主体的数据信息。然而即便是在原始数据获得良好保护的情况下，参与主体的私有数据仍存在泄漏的可能性。

模型梯度更新会导致隐私泄露. 尽管模型在训练的过程中已经使用了很多方法在防止原始数据泄露，在多方分布式的 AI 模型训练中，个体往往会使用自己的数据对当前的模型进行训练，并将模型的参数更新传递给其它个体或者中心服务器。在最近机器学习和信息安全的国际会议上，研究人员提出了一些利用模型参数更新来获取他人训练数据信息的攻击研究。Melis 等人 [36] 利用训练过程中其它用户更新的模型参数作为输入特征，训练攻击模型，用于推测其它用户数据集的相关属性；[37, 38] 等人利用对抗生成网络生成恢复其它用户的训练数据，在多方协作训练过程中，利用公共模型作为判别器，将模型参数更新作为输入数据训练生成器，最终可以获取受害者特定类别的训练数据。在最近的一项工作中 [39]，研究人员并未使用 GAN 等生成模型，而是基于优化算法对模拟图片的像素进行调整，使得其在公共模型上反向传播得到的梯度和真实梯度相近，经过多轮的优化模拟图片会慢慢接近真实的训练数据。

3.3 AI 系统安全性问题

AI 系统安全性问题与传统计算机安全领域中的安全问题相似，威胁着 AI 技术的保密性、完整性和可用性。AI 系统安全问题主要分为两类：1) 硬件设备安全问题，主要指数据采集存储、信息处理、应用运行相关的计算机硬件设备被攻击者攻击破解，例如芯片、存储媒介等；2) 系统与软件安全问题，主要指承载 AI 技术的各类计算机软件中存在的漏洞和缺陷，例如：承载技术的操作系统、软件框架和第三方库等。

3.3.1 硬件设备安全问题

硬件设备安全问题指 AI 技术当中使用的基础物理设备被恶意攻击导致的安全问题。物理设备是 AI 技术构建的基础，包含了中心计算设备、数据采集设备等基础设施。攻击者一旦能够直接接触相应的硬件设备，就能够伪造和窃取数据，破坏整个系统的完整性。例如：劫持数据采集设备，攻击者可以通过 root 等方式取得手机摄像头的控制权限，当手机应用调用摄像头的时候，攻击者可以直接将虚假的图片或视频注入相关应用，此时手机应用采集到的并不是真实的画面，使人工智能系统被欺骗；侧信道攻击，指的是针对加密电子设备在运行过程中的时间消耗、功率消耗或电磁辐射之类的侧信道信息泄露而对加密设备进行攻击的方法，这种攻击可以被用来窃取运行在服务器上的 AI 模型信息 [54]。

3.3.2 系统与软件安全问题

系统与软件安全问题是指承载 AI 应用的各类系统软件漏洞导致的安全问题。AI 技术从算法到实现是存在距离的，在算法层面上开发人员更关注如何提升模型本身性能和鲁棒性。然而强健的算法不代表着 AI 应用安全无虞，在 AI 应用过程中同样会面临软件层面的安全漏洞威胁，如果忽略了这些漏洞，则可能会导致关键数据篡改、模型误判、系统崩溃或被劫持控制流等严重后果。

以机器学习框架为例，开发人员可以通过 Tensorflow、PyTorch 等机器学习软件框架直接构建 AI 模型，并使用相应的接口对模型进行各种操作，无需关心 AI 模型的实现细节。然而不能忽略的是，机器学习框架掩盖了 AI 技术实现的底层复杂结构，机器学习框架是建立在众多的基础库和组件之上的，例如 Tensorflow、Caffe、PyTorch 等框需要依赖 Numpy、libopencv、librosa 等数十个第三方动态库或 Python 模块。这些组件之间存在着复杂的依赖关系。框架中任意一个依赖组件存在的安全漏洞，都会威胁到整个框架以及其所支撑的应用系统。

研究表明在这些深度学习框架及其依赖库中存在的软件漏洞几乎包含了所有常见的类型，如堆溢出、释放对象后引用、内存访问越界、整数溢出、除零异常等漏洞，这些潜在的危害会导致深度学习应用受到拒绝服务、控制流劫持、数据篡改等恶意攻击的影响 [15]。例如：360 Team SeriOus 团队曾发现由于 Numpy 库中某个模块没有对输入进行严格检查，特定的输入样本会导致程序对空列表的使用，最后令程序陷入无限循环，引起拒绝服务的问题。而在使用 Caffe 依赖的 libjasper 视觉库进行图像识别处理时，某些畸形的图片输入可能会引起内存越界，并导致程序崩溃或者关键数据（如参数、标签等）篡改等问题 [82]。另外，由于 GPU 设备缺乏安全保护措施，拷贝数据到显存和 GPU 上的运算均不做越界检查，使用的显存在运行结束后仍然存在，这都需要用户手动处理，如果程序中缺乏相关处理的措施，则可能存在内存溢出的风险 [83]。

第四章　AI 威胁常用防御技术

在第三章中，我们系统性地总结了 AI 模型、AI 数据以及 AI 承载系统面临的威胁。AI 模型面临的威胁包括：训练阶段的投毒与后门攻击、测试阶段的对抗攻击以及 AI 模型本身存在的鲁棒性缺失问题；AI 数据面临的威胁包括：利用模型查询结果的模型逆向攻击、成员推断攻击和模型窃取攻击，以及在训练阶段利用模型参数更新进行的训练数据窃取攻击；AI 承载系统面临的威胁包括：软件漏洞威胁和硬件设备安全问题等。

在本章节，我们将逐一介绍应对上述 AI 模型、AI 数据以及 AI 承载系统威胁的防御方法，并根据这些防御方法的特点进行相应的分类，总结不同防御方法的主要思想和针对的攻击场景，使读者对 AI 安全中的防御技术有初步而全面的了解。

表 4.1: 防御方法概括

威胁类型	防御种类	防御方法	针对具体攻击	攻击对应章节	代表性论文	本文介绍章节
数据投毒威胁	面向训练数据的防御	频谱分析法	标签翻转攻击	3.1.1	[84]	4.1.1
		激活值聚类法	标签翻转攻击		[9]	
		强扰动输入	常规后门攻击		[85]	
	面向模型防御	网络裁剪法	常规后门攻击	3.1.1	[10]	4.1.2
		后门逆向法	基于图案触发器的攻击		[86] [87][88]	
		模式连通法	常规后门攻击		[89]	
		ULP	常规后门攻击		[90]	
对抗样本威胁	对抗训练	FGSM 对抗训练	FGSM 对抗攻击	3.1.2	[23]	4.1.3
		PGD 对抗训练	常规对抗攻击		[91]	
		集成对抗训练	黑盒对抗攻击		[92]	
		Logits 对抗训练	常规对抗攻击		[93]	
		生成对抗训练	常规对抗攻击		[94]	
	输入预处理防御	输入变换法	灰盒、黑盒攻击	3.1.2	[11] [95] [96]	4.1.4
		输入清理法	灰盒、黑盒攻击		[97] [98] [99] [100]	
	特异性防御算法	防御性蒸馏法	FGSM 对抗攻击	3.1.2	[101]	4.1.5
		特征剪枝法	黑盒攻击与常规攻击		[102]	
		随机算法	黑盒攻击与常规攻击		[103]	
数据隐私威胁	模型结构防御	模型泛化法	模型窃取、成员推断攻击	3.2.1	[104][12][77][105]	4.2.1
		目标优化法	成员推断攻击		[78]	
	信息混淆防御	截断混淆法	模型窃取、成员推断攻击	3.2.1, 3.2.2	[7][73] [12] [104]	4.2.2
		噪声混淆法	成员推断攻击		[75] [106] [107]	
	查询控制防御	样本特征检测法	模型窃取、成员推断攻击	3.2.1	[75] [108] [74]	4.2.3
		用户行为检测法	成员推断攻击		[107]	

4.1 AI 模型自身安全性增强

3.1章节介绍了 AI 模型自身在训练与测试阶段遇到的安全威胁，包括投毒攻击、对抗样本攻击和鲁棒性缺乏威胁。为了应对这些威胁，学术界与工业界已经提出了许多有效的防御方法。这些防御方法从模型自身性质出发，针对性地增强了模型自身在真实场景下的鲁棒性。

AI 模型训练阶段主要存在的威胁是数据投毒攻击，它可以非常隐蔽地破坏模型的完整性。近些年来，研究者们提出了多种针对数据投毒攻击的防御方法。由于传统意义上的有目标的数据投毒攻击可以看作是后门攻击的一种特殊情况，因此后续章节将主要阐述针对后门攻击的防御方法。根据防御技术的部署场景，这些方法可以分为两类，分别是面向训练数据的防御和面向模型的防御。面向训练数据的防御部署在模型训练数据集上，适用于训练数据的来源不被信任的场景；面向模型的防御主要应用于检测预训练模型是否被毒化，若被毒化则尝试修复模型中被毒化的部分，这适用于模型中可能已经存在投毒攻击的场景。

AI 模型在预测阶段主要存在的威胁为对抗样本攻击。近些年来，研究者们提出了多种对抗样本防御技术，这些技术被称为对抗防御 (Adversarial Defense)。对抗防御可以分为启发式防御和可证明式防御两类。启发式防御算法对一些特定的对抗攻击具有良好的防御性能，但其防御性能没有理论性的保障，意味着启发式防御技术在未来很有可能被击破。可证明式防御通过理论证明，计算出特定对抗攻击下模型的最低准确度，即在理论上保证模型面对攻击时性能的下界。但目前的可证明式防御方法很难在大规模数据集上应用，我们将其作为模型安全性测试的一部分放在之后的章节阐述。本节主要阐述部分具有代表性的启发式防御技术，根据防御算法的作用目标不同分为三类：分别是对抗训练、输入预处理以及特异性防御算法。对抗训练通过将对抗样本纳入训练阶段来提高深度学习网络主动防御对抗样本的能力；输入预处理技术通过对输入数据进行恰当的预处理，消除输入数据中可能的对抗性扰动，从而达到净化输入数据的功能；其他特异性防御算法通过修改现有的网络结构或算法来达到防御对抗攻击的目的。

除了训练与预测阶段存在的威胁，AI 模型还存在鲁棒性缺乏风险。鲁棒性缺乏是指模型在面对多变的真实场景时泛化能力有限，导致模型产生不可预测的误判行为。为了增强 AI 模型的鲁棒性，提高模型的泛化能力，增强现实场景下模型应对多变环境因素时模型的稳定性，研究人员提出了数据增强和可解释性增强技术：数据增强技术的目标是加强数据的收集力度并增强训练数据中环境因素的多样性，使模型能够尽可能多地学习到各种真实场景下的样本特征，进而增强模型对多变环境的适应性；可解释性增强技术的目标是解释模型是如何进行决策的以及为何模型能够拥有较好的性能。若能较好地解答上述问题，将有助于在 AI 模型构建过程中依据可解释性的指导，有针对性地对模型进行调整，从而增强其泛化能力。

下面，我们将逐一介绍不同类别防御技术的主要思想和研究实例。

4.1.1　面向训练数据的防御

面向训练数据的防御试图保护模型在使用不信任来源的数据训练后，不受到后门攻击的影响。Tran 等人在文献 [84] 中考虑到：训练集中如果同时含有干净数据和毒化数据，毒化数据中的后门会在分类过程中提供一个很强的信号，只要这个信号足够大，就可以使用频谱特征（Spectral Signatures）进行奇异值分解来区分毒化数据和干净数据。防御者首先使用含有后门的数据集 D_{train} 训练得到模型 θ_p，提供特征表示 $\mathcal{R}$，并对每一个类别 y 中的所有数据计算特征表示的期望 $\widehat{\mathcal{R}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{R}(x_i)$ $(n = |D_y|)$。随后，计算标准化后的特征矩阵 $M = \left[\mathcal{R}(x_i) - \widehat{\mathcal{R}}\right]_{i=1}^{n}$ 的最大奇异值向量 v 以及每一个数据的异常值得分 $\tau_i = \left(\left(\mathcal{R}(x_i) - \widehat{\mathcal{R}}\right)\cdot v\right)^2$。最后，从 D_y 中去除异常值得分前 $1.5\cdot\epsilon$ 的数据并重新训练得到模型 θ，其中参数 ϵ 表示含有后门的毒化数据占全部数据比例的上界。作者从鲁棒统计的角度证明了这种方法的有效性。

Chen 等人在文献 [9] 中提出基于激活值聚类（Activation Clustering）的方法来检测含有后门的数据。他们认为含有后门的任意类别样本与不含后门的目标类别样本若能得到相同的分类结果，会在神经网络的激活值中体现出差异。在使用收集的数据训练得到模型后，他们将数据输入到模型并提取模型最后一层的激活值，然后使用独立成分分析（Independent Component Analysis, ICA）将激活值进行降维，最后使用聚类算法来区分含有后门的数据和正常的数据。

Gao 等人在文献 [85] 提出 STRIP 算法来检测输入数据中是否含有后门。他们对输入数据进行有意图的强扰动（将输入的数据进行叠加），利用含有后门的任意输入都会被分类为目标类别的特点（若模型含有后门，含有后门的输入数据在叠加后都会被分类为目标类别，而正常数据叠加后的分类结果则相对随机），通过判断模型输出分类结果的信息熵来区分含有后门的输入数据。

4.1.2　面向模型的防御

面向模型的防御试图检测模型中是否含有后门，若含有则将后门消除。Liu 等人在 [10] 中提出使用剪枝 (Pruning)、微调 (Fine Tuning) 以及基于微调的剪枝 (Fine Pruning) 等三种方法来消除模型的后门。他们基于 Gu 等人在 [2] 中发现的后门触发器会在模型的神经元上产生较大的激活值使得模型误分类的现象，提出通过剪枝的操作来删除模型中与正常分类无关的神经元的方法来防御后门攻击。他们提取正常数据在模型神经元上的激活值，根据从小到大的顺序对神经网络进行剪枝，直到剪枝后的模型在数据集上的正确率不高于预先设定的阈值为止。然而，若攻击者意识到防御者可能采取剪枝防御操作，将后门特征嵌入到与正常特征激活的相关神经元上，这种防御策略将会失效。应对这种攻击，研究人员发现通过使用干净数据集对

模型进行微调便可以有效地消除模型中的后门，因此结合剪枝和微调的防御方法能在多种场景下消除模型中的后门。

Wang 等人在文献 [86] 中提出 Neural Cleanse 来发现并消除模型中可能存在的后门。他们根据含有后门的任意输入都会被分类为目标类别的特点，通过最小化使得所有输入都被误分类为目标类别的扰动来逆向模型中存在的后门，其过程可以总结为如下公式：

$$\min_{m,\Delta} \ell\left(y_t, f(A(x), m, \Delta)\right) + \lambda \cdot |m| \quad \text{for } x \in X \tag{4.1}$$

其中，A 表示公式3.6中给数据 x 添加后门触发器的操作，f 表示模型，y_t 表示攻击者的目标类别。为了消除后门，他们将逆向得到的后门触发器添加在数据上，并将这些数据的类别标记正确，最后使用这些数据对后门模型进行训练。

然而 Neural Cleanse 需要使用一个干净的数据集对模型进行白盒查询来逆向后门，并且上述的后门逆向操作需要对每一个类别进行，这导致该方法无法应用到类别较多的模型中。因此，Chen 等人在 [87] 中提出在黑盒场景下（模型和训练数据未知）发现并消除模型中后门的方法：DeepInspect。他们首先通过模型逆向 [6] 得到替代的训练数据集，然后通过条件对抗生成网络 cGAN 来模拟可能存在的后门触发器的分布。cGAN 中的生成器可以用 $G(\mathbf{z}, t)$ 表示，其中 $\mathbf{z}$ 为随机噪声向量，t 为目标类别，黑盒查询模型则作为判别器 D。根据得到的后门触发器分布，他们使用基于假设检验的异常检测技术来判断输入样本中是否含有后门，最后可以通过类似 [86] 的方法消除模型中可能存在的后门。

Qiao 等人在 [109] 中发现 [86] 中的逆向方法在不同随机种子下会得到不同的后门触发器，甚至有些后门触发器的攻击效果要比原本的后门触发器好，因此 [86] 中的防御方法在某些情况下效果并不好。基于上述的观察，他们提出由一组生成模型集成得到的最大熵阶梯近似器 (Max-Entropy Staircase Approximator, MESA) 来近似模型中后门的分布。MESA 采用阶梯近似使用一组生成模型来学习后门的分布，每一种子生成模型只需要学习部分分布，这降低了对高维度数据建模的复杂度。MESA 中的每一种子模型都是基于最大化熵 (Entropy Maximization) 生成的，这可以避免直接从原始的后门触发器分布中进行采样，因为防御者无法知道后门触发器的分布。他们将最大化熵转化为最大化互信息 $I(X|Z)$，其中 X 为输入噪声，Z 为生成模型的输出，这种方法能够缓解 GAN 等生成模型中的模式崩塌 (Mode Collapse) 问题。MESA 采用 [110] 中提出的基于神经网络的互信息估计 (Mutual Information Neural Estimation, MINE) 来求解上述问题。最后，他们根据 MESA 生成的后门分布对含有后门的模型进行重新训练来消除其中的后门。

Liu 等人在 [88] 提出一种可以扫描并消除模型中可能存在后门的系统：ABS。由于后门攻击的本质是利用模型的冗余性来嵌入攻击者指定的后门特征，因此他们通过对模型神经元的激活值进行不同程度的修改，观察模型输出的差异，这样便可以

找到后门特征所嵌入的神经元从而逆向得到攻击者嵌入的后门，并通过对含有后门的模型进行重新训练，消除其中的后门。

Kolouri 等人在 [90] 中提出使用通用快速测试后门的图案（Universal Litmus Patterns, ULPs），检测模型中是否含有后门。他们事先生成大量正常和含有后门的模型，进行监督学习同时优化得到 ULPs，以及根据 ULPs 来判断模型中是否含有后门。在检测过程中，将 ULPs 输入给未知模型，并把得到的输出输入给二分类器来判断该未知模型中是否含有后门。值得注意的是，这种方法虽然能够快速检测出模型中是否含有后门，但是却不能修复其中的后门特征。

除此之外，Zhao 等人在 [89] 中使用基于模式连通性 (Mode Connectivity) [111] 的方法在仅借助少量干净数据的情况下来修复两个含有后门的模型。对于具有相同架构且经同样的损失函数训练得到的两个模型，模式连通性表明这两个模型可以在损失空间上通过一条高精准度（低损失）的简单曲线进行连接。他们从数据和模型空间的相似性比较说明了该防御的有效性。

4.1.3　对抗训练

对抗训练 (Adversarial Training) 是针对对抗攻击的最为直观防御方法，它使用对抗样本和良性样本同时作为训练数据对神经网络进行对抗训练，训练获得的 AI 模型可以主动防御对抗攻击。对抗训练过程可以被归纳为 MIN-MAX 过程，表述为

$$\min_{\theta} \max_{D(x,x')<\alpha} J(\theta, x', y) \tag{4.2}$$

其中 $J(\theta, x', y)$ 是对抗攻击的损失函数，θ 是模型参数，x' 是对抗样本，y 是样本 x 的正确标签。内部的最大化损失函数的目的是找出有效的对抗样本，外部的最小化优化问题目的是减小对抗样本造成的损失函数升高。

FGSM 对抗训练 (FGSM Adversarial Training) 首先由 Goodfellow 等人在文章 [23] 中提出。他们使用良性样本和 FGSM 算法生成的对抗样本一起训练神经网络，损失函数可以表述为：

$$J(\theta, x, y) = cJ(\theta, x, y) + (1-c)J(\theta, x + \epsilon \cdot sign(\bigtriangledown_x J(\theta, x, y)), y) \tag{4.3}$$

其中，$x + \epsilon \cdot sign(\bigtriangledown_x J(\theta, x, y)), y)$ 是良性样本 x 通过 FGSM 算法生成的对抗样本。文章 [23] 的实验结果表明，FGSM 对抗训练可以使 FGSM 对抗样本的攻击成功率由 89.4% 下降为 17.4%。尽管 FGSM 对抗训练对 FGSM 攻击有效，但是训练后的模型仍然容易受到基于迭代方法生成的对抗样本攻击。

针对基于迭代方法的对抗攻击，Madry 等人在文章 [91] 中提出 PGD 对抗训练 (PGD Adversarial Training)，PGD 对抗训练使用良性样本和由 PGD 算法生成的对抗样本一起训练神经网络。经过 PGD 对抗训练的神经网络能抵抗几种具有代表性的一阶 L_∞ 攻击，如在黑盒和白盒设置下的 FGSM、PGD 和 CW_∞ 攻击。实验结果

表明，PGD 对抗训练总体上是针对 L_∞ 攻击最有效的对策，然而生成 PGD 对抗样本需要的大量计算成本。

为了解决对抗训练模型易受黑盒攻击的问题，文章 [92] 提出了集成对抗训练算法 (Ensemble Adversarial Training)。该算法预先训练多个具有不同网络结构的模型，然后针对这些不同的模型分别生成对抗样本，将它们用于对抗训练。这种方法增加了用于训练的样本多样性，从而可以防御对抗样本的迁移攻击。实验结果表明，集成对抗训练获得的模型对单步和多步攻击产生的对抗样本迁移攻击都具有较强的防御性能。

文章 [93] 提出基于 Logit 配对的对抗训练算法 (Adversarial Logit-pairing, ALP)。该算法通过修改损失函数，在训练过程中激励神经网络在良性样本和其对抗样本在 Logit 层具有相似的输出，从而使神经网络可以对于对抗样本给出与之相对应的良性样本一致的判断。训练过程中使用的损失函数是：

$$J(\theta, x, x', y) = J(\theta, x, y) + cJ(\theta, x, x') \tag{4.4}$$

其中，$J(\theta, x, y)$ 是原始损失；$J(\theta, x, x')$ 是原样本 x 和对应的对抗样本 x' 的 Logit 层的交叉熵。文章 [93] 中的实验表明，这种 Logit 配对损失有助于在多个基准数据集上提高 PGD 对抗训练的性能，如 SVHN、CIFAR-10 和小型 ImageNet。

上述对抗训练算法均采用特定的攻击算法生成对抗样本，并将它们纳入训练过程。与之不同，文章 [94] 提出生成对抗训练。他们使用辅助分类生成对抗网络（AC-GAN）训练一个对抗样本生成器，该生成器可以生成带有标签的行为类似 PGD 对抗样本的训练数据。这些生成器生成的对抗样本将与良性样本一起用于训练鲁棒的分类器。

4.1.4　输入预处理防御

基于输入预处理的对抗防御方法通过对输入数据进行恰当的预处理，消除输入数据中存在的对抗性扰动。预处理后的输入数据将代替原输入样本输入网络进行分类，使模型获得正确的分类结果。输入预处理防御是一种简单有效的防御方法，它可以很容易地集成到已有的 AI 系统中。例如：图片分类系统中的预处理模块与分类模型通常是解耦的，因此很容易将输入预处理防御方法集成到预处理模块中。

一类数据预处理方法使用 JPEG 压缩、滤波、图像模糊、分辨率调整等方来对输入图像进行预处理。例如, Xie 等人在文章人 [11] 中提出一种将图片压缩到随机大小，然后再将压缩后的图片固定到随机位置并向周围补零填充的预处理防御方法。他们的这种防御方式在 NIPS 2017 对抗防御比赛黑盒赛道中获得了较好的成绩，但在白盒攻击情况下可被 EoT 算法攻击成功。Guo 等人 [95] 尝试使用位深度减小、JPEG 压缩、总方差最小化和图像缝合等操作对输入样本进行输入预处理，以减轻对抗性扰动对模型分类结果的影响。这种方法可以抵抗多种主流攻击方法生成的灰盒和黑

盒对抗样本，但是它仍易受 EoT 算法的攻击。输入预处理除了可以直接防御对抗攻击，还可以实现对抗样本的检测。例如，Xu 等人 [96] 分别利用位深度减小和模糊图像两种压缩方法对输入图像进行预处理，以减少输入样本的自由度，从而消除对抗性扰动。他们通过比较原始图像和被压缩图像输入模型后的预测结果的差异大小，辨别输入数据是否为对抗样本。如果两种输入的预测结果差异超过某一阈值，则将原始输入判别为对抗样本。

另一类输入预处理技术依赖于输入清理 (Input Cleansing) 技术。与传统的基于输入变换的输入预处理技术不同，输入清理利用机器学习算法学习良性样本的数据分布，利用良性样本的数据分布精准地去除输入输入样本中的对抗性扰动。研究人员首先尝试基于生成对抗网络 (GAN) 学习良性样本的数据分布，并使用 GAN 进行输入清理。Samangouei 等人 [97] 提出使用防御对抗生成网络 (Defense-GAN) 进行输入清理。他们训练了一个学习了良性样本数据分布的生成器，输入数据在输入神经网络进行分类前，会预先在 Defense-GAN 学习到的数据分布中搜索最接近于原输入数据的良性样本。该良性样本将替代原样本，输入神经网络进行预测，使模型输出正确的预测结果。Shen 等人 [98] 提出使用扰动消除对抗生成网络 (Adversarial Perturbation Elimination GAN，APE-GAN) 进行输入清理，其生成器的输入是可能为对抗样本的源数据，输出是清除对抗性扰动后的良性样本。此外，自动编码器技术也被证明可用于输入清理。Meng 等人 [99] 使用良性样本数据集训练具有良性样本数据分布的自动编码器。该自动编码器将把可能作为对抗样本的输入进行重构，输出与之对应的良性样本。与上述工作仅从输入层面进行输入清理不同，Liao 等人 [100] 尝试从更深层次的深度学习网络特征图层面进行输入清理。为了使输入样本在特征层面上没有对抗性扰动，他们提出了高阶表征指导的去噪器 (High-level Representation Guided Denoiser, HGD)。HGD 训练了采用特征级损失函数的用于降噪的 U-Net，最大程度地减少良性样本和对抗样本在高维特征中的差异，从而去除对抗扰动。

基于输入预处理的防御把防御的重点放在样本输入网络之前，通过输入变换或输入清理技术消除了对抗样本中的对抗性扰动，处理后的样本对模型的攻击性将大幅减弱。输入预处理防御可以有效地防御黑盒和灰盒攻击，然而对于在算法和模型全部暴露给攻击者的白盒攻击设置下，这些算法并不能保证良好的防御性能。

4.1.5 特异性防御算法

除了对抗训练和输入预处理，很多工作通过优化深度学习模型的结构或算法来防御对抗攻击，我们将其称之为特异性防御算法。进年来，越来越多的启发式特异性对抗防御算法被提出，我们选取其中一些具有代表性的算法归纳如下。

蒸馏算法被证明可以一定程度提高深度学习模型的鲁棒性。Hinton 等人 [112] 最早提出蒸馏算法 (Distillation)，该算法可以做到从复杂网络到简单网络的知识迁移。Papernot 等人 [101] 在此基础上进一步提出了防御性蒸馏算法 (Defensive Distil-

lation)。防御性蒸馏模型的训练数据没有使用硬判别标签 (明确具体类别的独热编码向量)，而是使用代表各类别概率的向量作为标签，这些概率标签可由早期使用硬判别标签训练的网络获得。研究者发现防御性蒸馏模型的输出结果比较平滑，基于优化的对抗攻击算法在攻击这种模型较难获取有效的梯度，因此防御性蒸馏网络获得了较好的对抗攻击的鲁棒性。实验结果显示，防御性蒸馏算法可以有效地防御 FGSM 对抗攻击。

对深度学习网络进行特征修剪可以有效地防御传统的对抗攻击手段。例如，Dhillon 等人 [102] 提出随机特征剪枝算法 (Random Feature-Pruning, SAP)。该算法随机修剪每一特征层中部分被激活的特征，其中特征值大的激活项具有更大的概率被保留，特征剪枝后会根据被保留的特征重新对该特征层进行正则化操作。

向深度学习网络中加入随机化操作可以有效防御黑盒模型下的对抗攻击。例如，Liu 等人 [103] 提出向深度学习网络中加入噪音层来减弱对抗性扰动特征的影响，并将其命名为随机自集成算法 (Random Self-Ensemble，RSE)。RSE 算法通过向卷积层前添加高斯噪音层，预防基于梯度的扰动攻击，并在模型预测阶段集成多次的预测结果作为最终的分类结果，以稳定模型的性能。向网络中加入随机的高斯噪音层对黑盒攻击以及常规的对抗攻击具有较好的防御效果，但被 Athalye 等人 [52] 证明这类基于随机的对抗防御算法均可被 EoT 算法攻击。

4.1.6 鲁棒性增强

鲁棒性增强是指在复杂的真实场景下，增强 AI 模型面对环境干扰以及多样输入时的稳健性。目前，AI 模型仍然缺乏鲁棒性，当处于复杂恶劣的环境条件或面对非正常输入时，性能会出现一定的损失，做出的不尽人意的决策。鲁棒性增强就是为了使模型在上述情况下依然能够维持其性能水平，减少意外的决策失误，可靠地履行其功能。构建高鲁棒性的 AI 模型不仅有助于提升模型在实际使用过程中的可靠性，同时能够从根本上完善模型攻防机理的理论研究，是 AI 模型安全研究中重要的一部分。为了增强模型的鲁棒性，可以从数据增强和可解释性增强两个方面进行深入探索。

数据增强

数据增强是指对当前拥有的数据集进行适当的修改，提高数据的数量和质量，从而实现模拟真实环境，提高模型的泛化能力。当今 AI 技术仍然处于大规模数据驱动的学习阶段，训练数据的质量是决定模型性能的关键因素之一。然而这些训练集仍然不能涵盖现实场景中出现的所有情况，导致模型可能出现预料之外的结果。因此通过数据增强对已有的训练集进行大量的扩充是提高模型鲁棒性和泛化能力的重要手段之一。

以计算机视觉领域为例，我们可以通过对图像进行适当调整模拟现实场景下的环境因素进行数据增强，例如：对图片进行仿射变换、光照调节、翻转、裁剪、注入噪声、随机擦除或滤波等 [113]。除了基于模拟环境因素的数据增强，还可以基于深度学习算法进行数据扩充，例如：对抗训练，利用对抗攻击算法生成对抗样本对数据集进行补充，弥补模型薄弱的部分，从而增强模型面对恶意攻击时的鲁棒性 [91]；对抗生成网络，对抗生成网络（GAN）是 Goodfellow 等人 [114] 提出的一种生成模型，由一个生成器和辨别器组成。生成器的目标是接收噪音向量作为输入，尽可能地生成与已有数据相似的样本。辨别器则负责鉴别样本是真实的还是伪造的。通过这样对抗的训练方式，生成器可以很好地模拟训练集的数据分布生成逼真的样本，这样通过对抗生成网络就可以对缺少的数据集进行补充。

4.1.7　可解释性增强

可解释性增强一方面从机器学习理论的角度出发，在模型的训练阶段，通过选取或设计本身具有可解释性的模型，为模型提高性能、增强泛化能力和鲁棒性保驾护航；另一方面要求研究人员能够解释模型有效性，即在不改变模型本身的情况下探索模型是如何根据样本输入进行决策的。针对模型可解释性增强，目前国内外研究主要分为两种类型 [115, 116, 117]：**集成解释**（Integrated Interpretability）和**后期解释** (Post Hoc Interpretability)。

集成解释

集成解释是指在模型的训练阶段，通过选取和设计本身具有可解释性的模型来增强模型的可解释性，其主要目的在于研究模型是如何进行学习表达。神经网络可解释性的增强，有助于帮助模型获得更好的泛化能力，启发更多原理清晰和运行可靠的模型设计方法。根据设计目标的不同，集成解释可以分为**功能性增强**和**泛化能力增强**。

功能性增强，是指对模型进行设计，使模型的决策过程、模块功能以及特征学习等部分更加令人易于理解。

- **模拟性.** 如果一个模型的决策过程可以被人们完整地模拟演算出来，那么该模型具有模拟性，同时也具有可解释性。例如：在数据的特征数量较少的情况下，可以选择决策树（Decision Tree）[118] 或者规则列表 [119, 120] 等具有较高模拟性的模型。

- **模块化.** 如果一个模型各个功能部件相互独立，每一个部件的解释也相互独立，那么该模型是模块化的。选择这种模型可以提高可解释性。例如：深度学习里的 Attention 技术 [121] 和模块化网络结构 [122] 可以单独作为一个部件帮助分

析神经网络的内部运行情况。概率模型（Probabilistic Model）[123] 由于其条件化的独立结构而天然具有模块化属性，因此也可以用来解释模型的不同部件。

- **特征工程.** 特征工程是一种从数据中提取重要特征的数据分析技术 [124]，可以帮助模型更加有效地处理数据，同时也可以帮助人们理解模型决策过程中依靠的特征是什么。例如：在自然语言处理中，文档类数据可以通过 tf-idf[125] 技术编码成向量形式来提取重要特征。在自动化特征工程技术中，无监督学习和降维是两种代表性研究方向。无监督学习可以处理无标签数据并提取对数据的结构化描述，相关技术包括聚类（Clustering）[126]、矩阵分解（Matrix Factorization）[127] 等。降维技术可以把高维数据转化为一种低维的表示形式，旨在提取更为关键的特征，相关方法包括主成分分析（Principle Component Analysis）等。

- **泛化能力增强.** 泛化能力是衡量 AI 模型性能的重要指标，用于评判 AI 模型在训练数据集之外的真实数据集上的预测效果。研究人员在传统的机器学习研究中发现，过高的模型复杂度可能会引起模型过拟合等问题，导致泛化误差增大。因此我们经常限制模型的非零参数数量和显式正则化等手段降低模型复杂度，使模型工作过程容易令人理解。例如，在损失函数中添加惩罚项 [128]。对于深度学习，模型的参数远多于其训练过程使用的训练样本数量。在大量的实验中，研究者们发现，当模型参数增加，变得越来越复杂时，模型可以很快地去拟合有限的数据，而在测试集上的泛化误差并没有增大甚至还会继续略微下降，这种现象称为 Overparametrization。尽管研究表明显式的正则化以及一些训练过程中产生的隐性正则化对泛化误差的降低有一定的作用，但不能根本地解释为何过于复杂的深度神经网络即使在没有正则化的情况下，仍然能取得较好的泛化性能力 [129]。研究者们做了许多工作试图解释这种现象：1) 对模型复杂度以及泛化能力的重新思考。实验表明模型规模的大小并不适合用于直接度量模型的复杂度 [130]，传统用于度量模型复杂度的方法也很难解释大型神经网络的泛化能力，因此有大量研究关注探索如何重新度量模型的复杂度及其对应的泛化能力 [129, 131, 132, 133]；2) Overparamization 和优化过程带来的隐性影响。例如，Overparamization 使得基于梯度下降的优化算法能够可靠地达到损失函数的全局极小值 [134]。另外，一些研究者认为模型的学习算法隐含着最小化参数范数的约束 [135]。然而目前的大多数理论研究还限于规模较小的简单双层网络，虽然取得实质性突破仍然需要更多的深入研究，但是这些理论研究中的某些结论和原则，仍然对构建优秀性能的 AI 模型有着指导意义。

后期解释

后期解释在模型训练完成之后，尝试理解深度学习模型。即：在具体的问题背景下，尝试去剖析模型对于真实输入样本的解释和刻画，解释模型为什么做出相应的决策，而不是仅仅留给使用者一个简单的结果。许多代表性工作围绕预测过程中的数据特征对预测结果的影响展开，试图掌握在数据集层面上，寻找样本哪些部分的特征对模型决策起重要作用。这些样本具有高重要分数，或者说具有显著性。因此有很多工作致力于寻找样本中影响决策的显著性特征。

- **直接可解释性分析.** 直接可解释性分析利用模型显著性信息（例如：中间特征、梯度和模型参数等），结合可视化等技术辅助，直接验证显著特征。例如：在计算机视觉任务中，可以利用深层卷积网络层的特征进行反卷积对图像关键区域进行定位 [136]，或者对相应的目标函数（例如：某个神经元激活值、分类任务中的各类别 Confidence 等）进行反向传播，求出图片上对应像素的梯度。梯度信息在一定程度上能够表示各个特征对于当前目标值的显著性，也可被用于重建与当前目标值相关的显著性图像特征，结合可视化技术，使用者就能直观地看出深度学习模型在做出决策的时候主要关注的区域和特征 [137, 138, 139, 140, 141]。上述工作直接对数据特征进行显著性分析，存在一定的不足：它们无法分析特征之间的关联性对模型决策的影响，而特征之间的相互作用往往对于科学探索和假设验证具有重要作用。目前有一些工作尝试探索这个问题。例如：文章 [142] 通过对模型的权重矩阵进行分析，得知样本特征的相互作用产生于网络的隐藏神经元中；文章 [143] 根据输入特征对于模型预测结果的贡献进行层次聚类，不同的类别特征之间的组合关联则会对决策产生不同的影响，进而可以对特征之间的相互作用进行一定的解释。

- **非直接可解释性分析.** 另一部分工作为了研究图片中影响 AI 模型决策结果的显著性特征，会有针对性地对样本特征进行选取或者修改，而显著性特征的修改对于模型决策结果的影响较大。例如：防御方会有针对性地向图片中的不同区域添加干扰 [144, 145]、进行模糊化和遮盖 [146] 等处理，观察比较不同部位的干扰对最终结果的影响，进而找到最能够保留模型决策结果的区域，这个过程可以人为地指导进行，也可以交给计算机自主进行，从而能够自动快速地标识样本显著性区域 [147]。因此当构建 AI 模型的时候，可以根据模型所关注部位的特点，在具体的任务中进行相应的改动，从而增强模型对于关键特征的识别，进而提高可解释性和泛化能力。例如：通过有选择地遮盖面容不同的部位，最小化其在特征层与原始面容的差异，进而使模型学习到结构化的特征，对意外地遮挡扰动有更强的鲁棒性 [148]。

4.2 AI 数据安全与隐私泄漏防御

3.2章节介绍了 AI 模型在训练和测试过程中可能会造成的模型与数据隐私泄漏风险，包括训练阶段模型参数更新导致的训练数据信息泄漏、测试阶段模型返回查询结果造成的模型数据泄漏和数据隐私泄漏。为了减轻这些 AI 模型正常使用过程中间接引起的数据隐私泄漏，学术界和工业界从不同角度进行了许多尝试。

在之前3.2章节的介绍中，我们可以知道：即便在没有被直接攻击破解的情况下，AI 模型正常训练和使用的过程中产生的信息也会导致数据隐私的间接泄漏。为了解决这类数据隐私泄漏，研究人员们采用的主要思想就是在不影响 AI 模型有效性的情况下，尽可能减少或者混淆这类交互数据中包含的有效信息。AI 模型部署时可以采用以下几类数据隐私保护措施：模型结构防御，该类方法是指在模型的训练过程中对模型进行有目的性地调整，降低模型输出结果对于不同样本的敏感性；信息混淆防御，该类方法通过对模型输出和模型参数更新等交互数据进行一定的修改，在保证模型有效性的情况下，尽可能破坏混淆交互数据中包含的有效信息；查询控制防御，该类防御通过对查询操作进行检测，及时拒绝恶意的查询，从而防止数据泄漏。

4.2.1 模型结构防御

面向模型的防御是通过对模型结构做适当的修改，减少模型泄露的信息，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。

Fredrikson[104] 等人提出当目标模型为决策树时，可使用 CART 决策树的变种，将样本的敏感特征的优先级调高或调低。他们通过实验证明，当敏感特征在决策树的根节点和叶子节点层级时，对 Model Inversion 攻击能够达到较好的防御效果，其中当敏感属性位于根节点时，能达到最好的防御效果。

Shokri[12] 等人和 Ahmed[77] 等人提出可以在目标模型中添加 Dropout 层，或者使用 Model Stacking 的方法将不同的元学习器聚合在一起，又或者在目标模型中添加正则项。通过实验，他们发现当目标模型使用上述方法后，能显著地减少成员推断攻击的准确率。

Nasr[78] 提出了一种基于对抗学习的防御方法，他们认为如果能计算出当前模型抵抗成员推断攻击的成功率，并将其作为一个对抗正则项加入到损失函数中，那么在训练过程中使用 MIN-MAX 的对抗训练方式，最终就可以训练出一个有效抵抗成员推断攻击的模型。该模型受到的成员推断攻击的成功率将存在一个上界，我们可以使这个上界足够小的同时，让模型仍然保持有较高的分类准确度。

此外，Wang[105] 等人构建了 MIASec，他们提出可以对训练数据在目标模型的关键特征上进行特定的修改，使模型对成员数据和非成员数据的预测向量的分布难以区分，进而完成对成员推断攻击的防御。

如前文所说，模型逆向攻击的核心原因是输出向量包含了训练样本的信息，成

员推断攻击的核心原因是模型对训练样本和测试样本的预测向量的分布不一致。因此，防御模型逆向攻击就是尽可能地降低输出向量与输入向量之间的关联，防御成员推断攻击就是尽可能地缩小输出向量的分布差异。面向模型的防御旨在修改模型的结构和损失函数，使目标模型给出的输出向量中包含尽可能少的信息，从而取得较好的防御效果。但这种防御方式仍有缺陷，它会对目标模型的性能有较大影响，导致其分类准确度出现波动。因此，防御方需要在模型的性能与其鲁棒性之间做出平衡。

4.2.2 信息混淆防御

面向数据的防御是指对模型的预测结果做模糊操作。通过这些模糊操作，在保证 AI 模型输出结果正确性的前提下，尽可能地干扰输出结果中包含的有效信息，从而减少隐私信息的泄露。这些数据模糊操作主要包含两类：一类是截断混淆，即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 [7, 73]；另一类是噪声混淆，即对输出的概率向量中添加微小的噪声，从而干扰准确的信息 [75]。

针对于截断混淆，Shokri[12] 等人提出可以对目标模型的输出向量进行截取，比如只给出输出向量中前几高的类别和概率值，或者降低输出向量中小数位的保留位数；Fredrikson[104] 等人提出可以对目标模型的输出向量进行取整，达到对输出向量的修饰效果。通过截断混淆，我们能够防御模型逆向攻击和成员推断攻击。

针对于噪声混淆，Jia[106] 等人基于对抗样本的理念提出了 Memguard。他们发现成员推断攻击对目标模型给出的预测向量的变化非常敏感，如果我们为这些预测向量添加一个精心设计的噪声，从而混淆成员数据和非成员数据的预测向量分布的差异，就可以生成一个对实际结果没有影响的“对抗样本”，这样就可以完成对成员推断攻击的防御。He[107] 等人提出可以用差分隐私的方法对输出向量加噪声进行混淆。他们认为可以利用差分隐私的算法来移除输出向量自身的特征，但输出向量同时保留了其关于分类结果的信息，使其难以被区分。此外，他们还提出可以在损失函数中添加噪声项，在轻微地牺牲分类准确率的同时，提高输出向量的隐私性，完成对成员推断攻击的防御。

模型逆向攻击和成员推断攻击的输入都是目标模型的输出向量，因此，如果我们能够在不影响分类结果的前提下，对输出向量进行特定地修饰，就可以扰乱输出结果中的有效信息，从而完成防御。但这种防御方法依然具有局限性。如果对输出向量的修饰程度较小，则其抵抗攻击的能力也不会很好；如果对输出向量的修饰程度较大，则会影响分类数据的可用性。也就是说，这里仍然需要保证隐私性与可用性之间的平衡。

4.2.3 查询控制防御

查询控制防御是指防御方可以根据用户的查询行为进行特征提取，完成对隐私泄露攻击的防御。攻击者如果想要执行隐私泄露攻击，需要对目标模型发起大量的查询行为，甚至需要对自己的输入向量进行特定的修饰，从而加快隐私泄露攻击的实施。根据用户查询行为的特征，我们可以分辨出哪些用户是攻击者，进而对攻击者的查询行为进行限制或拒绝服务，以达到防御攻击的目的。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。

在异常样本检测中，攻击者为了窃取黑盒的在线模型，往往需要对在线模型进行大量的查询操作。出于提高窃取效率的目的，攻击者会对正常的样本进行有目的的修改。针对模型泄露攻击的特点，防御者主要通过检测对异常样本的查询，识别模型窃取行为。PRADA[75] 是一项针对模型窃取攻击进行检测的防御技术，它根据多个样本特征之间的距离分布来判断该用户是否正在施展模型窃取攻击。文章发现随机选取的正常样本特征间的距离大致服从正态分布，而模型窃取过程中查询的样本往往具有鲜明的人工修改迹象，其样本间距离分布也与正态分布区别较大。利用这一现象，防御方对若干次的查询进行统计检验则可检测出异常查询用户。此外，查询样本的特征分布也可以被用于检测。例如，Kesarwani[108] 等人记录下用户的查询的样本，检查其在特征空间中的分布，评估模型被盗取的风险；Yu[74] 等人提出正常样本的特征分布与人工修改的样本特征分布相比有较大的区别，防御方可以通过区分样本的特征分布来检测异常查询。

在查询行为检测中，由于攻击者往往需要对目标模型进行大量的查询，所以其查询行为与正常用户行为会有较大不同。根据这种差异，我们可以在一定程度上防御模型泄露和数据泄露攻击。针对数据泄露攻击的特点，He[107] 等人提出可以根据用户查询的行为特征，在样本输入阶段，完成对成员推断攻击的防御。攻击者实行成员推断攻击时有时需要大量查询目标模型，因此模型提供者可以根据用户的查询频率实现对查询次数的限制，从而提升攻击者部署成员推断攻击的成本。

由上文可知，防御方可以通过对异常样本的检测和异常查询行为的检测来完成对模型泄露攻击和数据泄露攻击的防御。但这种防御方法的针对性不强，效果不够好，误分类的概率较大。查询控制防御主要是在攻击模型的训练过程中起作用，对已训练好的攻击模型无能为力。此外，如果攻击者知道目标模型采用了查询控制防御，他们也有许多方法可以绕过这种防御方法，比如设计更难以被检测的异常样本，或者采用虚拟 IP 等方式绕过目标模型的检测。

4.3 AI 系统安全性防御

在3.3章节中我们提到，除了模型算法层面的威胁，AI 系统同样面临着来自硬件与软件层面上的安全问题。这些问题与传统计算机安全领域中的安全问题相似，威

胁着 AI 技术的保密性、完整性和可用性。为了保障 AI 技术能够安全稳定地落地应用，其硬件与软件安全同样不容忽视。目前研究者结合传统计算机安全保障技术，对 AI 系统的安全构建做出了许多实践探索。

硬件安全保护

- **设备加密.** 设备进行加密可以保障被加密的内部数据不被泄露。例如：华为手机在使用 AI 技术进行身份认证时，对于生物核身等隐私敏感数据，会通过安全通道将这些数据放入一个可信环境的安全隔离区中进行处理。从信息的采集、特征提取、特征比对到特征存储，这些敏感数据不会离开安全隔离区，外部也就无法获取内置安全芯片内的数据，这样确保了敏感数据不会泄露。

- **设备检测.** 对 AI 应用过程中使用的设备进行必要的检测，确保其不会被攻击者破坏劫持。这些设备包括手机、传感器等数据采集设备，以及服务器等计算资源设备。例如：某些特定的用户设备易于被恶意攻击者劫持，上传到服务器的数据是恶意篡改的数据。因此，防御方需要定期对 AI 应用中涉及的基础设施进行检查，拒绝为高风险设备持有者提供相关的服务，对重要设施实行严格的安全监管措施，避免其从物理层面和软件层面被攻击者劫持。

软件、系统安全保护

- **开源框架与及软件.** AI 应用的开发人员在使用开源框架以及软件的时候，应该详细阅读官方文档，严格遵守相应 API 的使用规范；在调用依赖包的时候，应对其版本与更新进行较为全面的了解，避免版本分歧细节导致的功能错误甚至程序崩溃；应该了解软件底层原理，避免在编写 AI 应用时造成算法范畴外的错误，增强代码的可扩展性可鲁棒性，保证 AI 应用安全。

- **权限分级管理.** 设置多级安全架构。对于各级使用人员进行严格的权限分级管理，根据职责授权，遵守数据可用不可见、任务与数据分离、授权进入的规章制度，保证执行授权边界清晰，保证系统安全；对于核心的模型数据进行加密，保证模型数据只能被可信任的程序访问调用。

- **操作行为可溯源.** 对于核心数据的活动，进行持续可追溯的管控措施，其生命周期内的操作要保留记录，生成记录事实和支持决策的审计跟踪、系统日志等；同时对于整个系统也要配备安全记录模块，将数据采集、输入样本、运行状态、系统输出等信息写入日志，方便在出现问题的时候回溯诊断追责。

第五章　AI 应用系统一站式安全解决方案

AI 技术已经是许多业务系统的核心驱动力，如苹果 Siri、微软小冰都依赖智能语音识别模型，谷歌照片利用图像识别技术快速识别图像中的人、动物、风景和地点。然而正如《人工智能安全》[149] 一书中提到，新技术必然会带来新的安全问题，一方面是其自身的脆弱性会导致新技术系统不稳定或者不安全的情况，这是新技术的内在安全问题，一方面是新技术会给其他领域带来新的问题，导致其他领域不安全，这是新技术的衍生安全问题。近年来学术界和工业界针对 AI 应用系统的攻击案例此起彼伏，例如腾讯攻破了特斯拉的自动驾驶系统、百度攻破了公有云上的图像识别系统、Facebook 和 Google 掀起了反 DeepFake 浪潮。

白皮书第 3 章介绍了 AI 系统是可能面临的包括对抗样本攻击、投毒攻击和供应链攻击等各类威胁，同时白皮书第 4 章也给出了面向各类 AI 威胁的防御技术。但在实际场景中，AI 系统遇到的威胁往往十分复杂，仅靠单一的防御技术无法有效抵御实际威胁。因此在本章节，我们先回顾国内外大厂采用的 AI 安全解决方案，然后再从这些方案中提炼出一套涵盖面更广泛的 AI 安全解决方案。

5.1　行业介绍

- **百度.** 百度是国内最早研究 AI 模型安全性问题的公司之一。当前百度建立了一套可衡量深度神经网络在物理世界中鲁棒性的标准化框架。事实上，物理世界中使用的模型往往与人们的衣食住行相关（如无人自动驾驶、医疗自动诊断等），这些模型一旦出现问题，后果将非常严重。因此，该框架首先基于现实世界的正常扰动定义了可能出现威胁的五大安全属性，分别是光照、空间变换、模糊、噪声和天气变化；然后，针对不同的模型任务场景，制定不同的评估标准，如非定向分类错误、目标类别错误分类到评估者设定的类别等标准；最后，对于不同安全属性扰动带来的威胁，该框架采用了图像领域中广为接受的最小扰动的 L_p 范数来量化威胁严重性以及模型鲁棒性。

- **腾讯.** 腾讯公司针对 AI 落地过程中面临的各类安全问题进行了细致的划分，

具体分为 AI 软硬件安全、AI 算法安全、模型安全、AI 数据安全和数据隐私等部分。软硬件安全主要是考虑到部署 AI 模型的软件和硬件层面可能存在的安全漏洞，如内存溢出、摄像头劫持等问题；AI 算法安全主要考虑深度学习存在对抗样本的问题，容易出现错误的预测结果；模型本身的安全则涉及到模型窃取，这一问题目前实现方式比较多，常见的方法是直接物理接触下载模型并逆向获取模型参数，以及通过多次查询来拟合“影子”模型实现等价窃取；此外，模型的训练数据也会被污染，开源的预训练模型可能被恶意埋入后门，这些问题都被划分为 AI 模型的数据安全问题；当然，模型训练使用的数据集也会涉及用户的隐私，因此攻击者可能也会通过查询获取用户隐私。为了缓解这些问题，腾讯安全团队借助 AI 能力，针对性地构建了多种攻击检测技术。

- **华为.** 华为公司同样对 AI 安全问题展开了深入的研究，其将 AI 系统面临的挑战分为 5 个部分，包括软硬件的安全、数据完整性、模型保密性、模型鲁棒性和数据隐私。其中，软硬件的安全涉及应用、模型、平台、芯片和编码中可能存在的漏洞或后门；数据完整性主要涉及各类数据投毒攻击；模型保密性则主要涉及到模型的窃取问题；模型鲁棒性考虑训练模型时的样本往往覆盖性不足，使得模型鲁棒性不强，同时模型面对恶意对抗样本攻击时，无法给出正确的判断结果等问题；数据隐私考虑在用户提供训练数据的场景下，攻击者能够通过反复查询训练好的模型获得用户的隐私信息。

  为了应对这些挑战，华为主要考虑三个层次的防御手段：攻防安全、模型安全和架构安全。其中，攻防安全考虑针对已知的攻击手段，设计针对性的防御机制来保护 AI 系统，经典的防御技术包括对抗训练、知识蒸馏、对抗样本检测、训练数据过滤、集成模型、模型剪枝等。而针对模型本身存在的安全问题，考虑包括模型可检测性、可验证性和可解释性等技术，以提升模型应对未知攻击的能力。在业务中实际使用 AI 模型，需要结合业务自身特点，分析判断 AI 模型架构安全，综合利用隔离、检测、熔断和冗余等安全机制设计 AI 安全架构与部署方案，增强业务产品、业务流程与业务功能的健壮性。

- **RealAI.** RealAI 是一家专注于从根本上增强 AI 的可靠性、可信性以及安全性的创业公司。该公司通过黑盒和白盒方式，对目标模型进行对抗样本攻击，并通过检测器和去噪器等方式构建模型的 AI 防火墙；此外，它们也考虑了模型窃取和后门检测等问题。

5.2　多维对抗与 AI SDL

AI 系统的防御与攻击者的攻击是一个不断演变的攻防对抗过程，攻击者会不断更新攻击手法来突破 AI 系统的防御。例如以黑产为代表的攻击者，会不断探测 AI

系统的漏洞，开发新的攻击工具，降低攻击成本来突破 AI 系统，获得高额的经济收益。

在实际场景中，我们需要从多个视角切入来应对与攻击者之间日益焦灼的对抗战役。一个非常有效的战略就是知己知彼，知彼就是从防御的视角切入，时时刻刻跟踪对手的动向，部署策略模型对各类攻击行为进行监测，对于这类技术我们称之为**多维对抗技术**，知己就是从评测的视角切入，实时检测 AI 系统中的漏洞并进行修补，降低攻击面、风险面，对于这类技术我们称之为 ***AI* 模型安全开发生命周（*AI SDL*）**，这也是借鉴应用安全领域的 SDL 理念。

多维对抗　多维对抗的核心理念就是把攻防链路进行切面（深度数据化），再充分融合机器智能和专家智能，结合威胁情报，化被动防御为主动攻防，在对手还在尝试阶段就能够发现异常行为，再通过置信度排序和团伙挖掘等进行审理定性、处置，是一个系统化的防御体系。

AI 模型安全开发生命周期（AI SDL）　*AI SDL* 是从安全角度指导 *AI* 模型开发过程的管理模式。*AI SDL* 是一个安全保证的过程，它在 *AI* 模型开发的所有阶段都引入了安全和隐私的原则。具体来说，*AI* 模型的生命周期包括模型设计、数据与预训练模型准备、模型开发与训练、模型验证与测试、模型部署与上线、模型性能监控、模型下线这七个流程。*AI SDL* 通过安全指导这 *7* 个模型开发流程，保障模型在其全生命周期中的安全性。

5.3　多维对抗

多维对抗从 AI 系统的多个切面数据出发，发现 AI 系统中的攻击样本与攻击手法。AI 系统中能有效反应攻击样本特征的切面数据有图像维度、账号维度、设备维度、用户行为维度等等，检测异常并分析这些异常是否为攻击样本。

5.3.1　多维异常检测

图像维度　在图像维度可以进行的检测有对抗样本检测、伪造样本检测、图像水印检测、图像指纹检测等。比如针对对抗样本检测，通过发现可疑的对抗样本，避免模型受到恶意攻击产生不可预计的决策。目前的检测手段主要通过特征区别和模型输出差异来分辨正常样本与对抗样本。文献 [150] 的作者提出收集良性样本和对抗样本组成训练集，抽取样本的 Local Intrinsic Dimensionality 特征用于训练一个分类器作为恶意样本检测器；文献 [96][99] 则分别使用自编码器、特征压缩等方式对输入样本进行变换，根据变换样本与原样本的预测差异进行分辨。比如针对伪造样本检测，开发者可以通过眨眼动作 [151]、色彩纹理特征 [152] 进行检测，另外大型数据集

[153][154] 的出现使得卷积神经网络、循环神经网络等深度学习技术成为了检测伪造攻击的有效手段 [155][156][157][158]。

行为维度　行为维度是分析攻击者的攻击前后的在 AI 系统上的用户行为，例如攻击者是否有薅羊毛行为、异常查看隐私数据行为和刷单行为等。如，当攻击者绕开 AI 生物核身等校验环节后，可能会利用虚假身份注册的用户来进行恶意活动以牟取非法利益。通过对后续行为序列等信息进行日志化，就可以构建对应的监测特征库，还可以使用支持向量机、逻辑回归、随机森林等有监督机器学习算法结合聚类等无监督方法对当前特征进行分析挖掘。

账号维度　账号维度可以进行的检测有垃圾账号检测、虚拟小号检测、虚假账号检测、僵尸账号检测等。安全技术人员可以对这些恶意账号的后续活动特征进行分析，进而检查出恶意账户以弥补 AI 应用自身的不足，可用于账号检测的信息包括账号之间的链接信息、账号的个体社会属性、账号活动行为以及用户行为的时序特征等。

设备维度　设备维度异常检测主要是分析用户设备是否存在异常使用，设备维度检测主要有恶意 APP 检测、设备指纹异常分析、设备伪造参数检测等。

5.4　AI SDL

AI SDL 是指导 AI 模型的安全开发，确保 AI 模型在模型设计、数据或预训练模型准备、模型开发训练、模型验证测试、模型部署上线、模型性能监控、模型下线这几个阶段的安全性。为了方便理解和应用，这 7 个 SDL 流程可以进一步简化为三大模块：模型评测与加固、业务评审和风险治理。

5.4.1　模型评测与加固

模型安全性评测

模型安全性评测是 AI SDL 中的一个重要的环节，模型上线前的安全性评测可以提前发现模型的漏洞并修复。在实际场景中，对 AI 模型有效的安全性评测有**模型可信评测**、**对抗鲁棒性评测**与**可解释性评测**。

- **模型可信评测**与传统软件工程中符号执行、模糊测试等技术思想类似，试图通过遍历模型决策空间所有的样本来对模型的行为进行测试，寻找出一些使模型做出异常判断的输入样本。目前开源的模型可信评测工具有 DeepXplore [159]、DLFuzz [160] 等；DeepXplore 通过神经元覆盖率来寻找异常样本；DLFuzz 通过神经元覆盖率与最大化变异样本与种子样本的欧式距离来寻找异常

样本；DeepGauge [161] 基于神经元、网络层不同粒度的覆盖率寻找异常样本；DeepHunter [161] 使用 6 种不同的覆盖率，通过变异亮度、对比度、缩放比例、旋转角度等寻找异常样本。这些模型可信评测思想均是从软件工程的角度出发，尽可能根据深度学习模型的特征，制定相应的测试策略自动生成大量测试样本对模型进行全面测试，从而找出潜在的会使模型产生异常的样本，对原来测试集进行有效的补充。

- **对抗鲁棒性评测**利用对抗攻击测试与数学验证等方法，测试 AI 模型在对抗攻击下的鲁棒性。对抗鲁棒性的测试主要分为攻击实验测试与可证明式鲁棒性测试：1) 攻击实验测试指使用对抗样本生成算法，生成攻击相应模型的对抗样本，通过实验测试其在攻击之下的性能。目前已经存在多种对抗攻防框架可供实现模型安全性评估 [162]：德国图宾根大学的研究人员创建的 Foolbox[163]、Nicolae 等人 [164] 构建的 Adversarial Robustness Toolbox 工具箱、Papernot 与 Goodfellow 等人持续维护的 Python 库 Cleverhans[165]、清华大学构建的 RealSafe 供测试平台等，利用这些工具和平台可以构建主流的对抗攻击算法，验证不同的场景设置下防御模型的对抗鲁棒性。

  2) 可证明式鲁棒性测试。给定待测试的模型，利用数学工具形式化地评估模型在一类攻击模式下的对抗鲁棒性。现有的可证明式鲁棒性测试主要以验证局部鲁棒性为主，即给定模型和相应的样本，形式化地验证在某个扰动范围内，是否存在对抗样本，或者模型最多能保持多少准确度。这些技术主要分如下四个类型：基于优化、基于可满足性求解器、基于区间分析和基于随机化的验证。基于优化的验证，利用了半正定松弛 [166][167]、对偶凸松弛 [168][169][170] 等数学优化方法来估计扰动输入域中最坏情况，从而可靠地估计模型在设置的攻击条件下，有多少样本可以保证不被攻击；基于求解器的验证，将验证问题转换为多个约束条件组成的一个约束系统，将模型在该样本处的鲁棒性验证转化为对约束系统下包含性问题的求解 [171][172][173][174]；基于区间分析的验证，这些工作会逐层对模型神经元的值域区间进行分析，判断某个样本在设置的扰动下最后结果层得到的输出是否可能导致误判 [175][176]；基于随机化的验证，Lecuyer 等人 [177] 从差分隐私的角度来重新审视对抗样本的问题，将样本扰动对模型输出的影响看作差分隐私问题，利用差分隐私条件量化地评估恶意扰动可能产生的影响，Cohen[178] 等人提出随机平滑（Random Smoothing）的机制，能够给出模型在 l_2 范数扰动下的可靠的鲁棒性边界，这种方法首次在 ImageNet 上取得的较好的可证明准确率，并且相对高效，可用于较大规模的具有更为普遍结构的模型中去。

- **可解释性测试**用来评估模型输出结果是否可靠、可依赖。可解释性的测试工作主要分为人工测试和自动化测试两类，以人工测试为主，自动化测试为辅。人

工测试的方法主要包括以应用为基础、以人为基础以及以功能为基础的方法[179]。以应用为基础主要是在一个真实应用场景中人工地做实验；以人为基础主要利用在简单任务上的人工评估结果；以功能为基础的方法不需要人类实验，但是需要一个定量的代理作为可解释性的保证。例如：在决策树模型中，可以使用树的深度作为这个定量的代理。文献 [180] 介绍了两种类型的可解释性：全局可解释性表示可以解释一个模型的全部情况；局部可解释性表示可以解释一个模型在特定输入和相应输出上的结果。这个方法邀请 1000 名用户根据输入数据的变化，写下他们期望的模型输出的变化，然后根据用户的结果跟模型真实的结果之间的匹配程度以及用户完成的时间，来估量模型的可解释性。

在自动化可解释性测试领域，文献 [181] 提出了一种度量方法来理解模型的行为。这种方法通过遮挡对象的周围环境来衡量模型是否在对象识别场景中学习了对象的方法。在书 [182] 中，作者提出了应该利用机器学习算法的类别来解释模型。作者认为，实现可解释性的最简单的方法就是使用那些本身可解释的模型，例如线性回归模型、逻辑回归模型和决策树模型。文献 [183] 定义了变质关系模式（Metamorphic Relation Patterns）和变质关系输入模式（Metamorphic Relation Input Patterns）两种概念来帮助用户理解机器学习系统是如何工作的。这个工作对很多种系统做了案例分析，包括大型商业网站、谷歌地图导航、基于谷歌地图定位的搜索、图片分析、视频分析等。机器学习技术广泛应用在医疗领域，但是模型预测结果缺乏医学上的解释。因此文献 [184] 提出把分类结果转换为疾病规模的可能性来提高模型的可解释性，并且表明，可以将任意等级的分类器评分校准为概率等级，而不会影响其分类性能。

AI 系统加固

数据加固　AI 模型训练数据的完整性、多样性、均衡性直接影响着模型的性能，但实际 AI 系统采集数据较差，存在着例如数据噪声、规模不足、数据集多样性较低不够均衡乃至数据投毒等多种问题。这样的数据会限制 AI 模型的能力和精度，甚至还有可能会隐藏巨大的安全隐患。因此需要对训练 AI 模型所使用的数据进行清洗、选择分析、扩充等预处理操作，使得数据满足数据完整性、多样性、均衡性。这一系列操作从训练数据层面加固了 AI 模型，又称之为**数据加固**。数据加固的目的是为了使模型训练数据满足完整性，即数据本身格式结构完整、语义信息清晰，相关标注准确无误，没有被破坏或篡改；多样性，即样本尽可能地涵盖多种类别；均衡性，即不同种类的数据不会出现过大的偏差，避免主观因素影响。

为了使数据能够满足以上的性质，数据加固主要采用**数据筛选**、**统计分析**、**数据预处理**等技术手法：

- **数据筛选**保证训练数据的质量，防止模型因为低质量数据而出现的性能损失甚至安全漏洞。从互联网或者用户群体中获取的大规模原始数据质量参差不一，

缺乏完整性，例如：从互联网中爬取的图像数据可能存在分辨率过低、干扰信息繁杂、语义信息不清甚至标注错误等多种质量问题，甚至存在恶意的毒化数据。如果 AI 模型使用的数据完整性受损，那么模型的性能就无法保证。

因此，在真实场景下使用数据训练模型之前要进行去除低质量数据、检查恶意数据等数据筛选操作，主要包括数据清洗和提纯等步骤。1) 数据清洗，检查语义信息与其标签不匹配的错误样本，修正标签或是放弃该样本，避免模型对错误数据的拟合从而保证性能和安全性，这个过程，除了人力检查，还可以训练一些简单的模型进行辅助以加快对海量数据的检查进程。为了避免遭受投毒攻击，还可以采用4.1.1中介绍的方法，过滤其中可能的毒化数据。2) 数据提纯，去除低质量样本，这些样本会对关键信息造成干扰，影响正常的训练过程，以图像任务为例，低质量数据包括损坏的不完整图像、低分辨率图像、语义信息模糊图像、干扰信息严重的图像等。

在不同领域的任务中，数据筛选的操作流程和具体标准都有所差异，但主要的目的就是通过修正或去除错误的低劣的数据来确保使用数据的完整性，提升数据质量以保证相关 AI 模型的性能和安全性。

- **统计分析**

  分布统计是对机器学习训练数据进行分析的常用技术，主要统计数据在不同标签或者某些属性上的分布情况。收集到的原始数据往往存在多种缺陷，这些缺陷可能会引起数据分布的异常，例如：不同类别或属性的数据量分布可能存在的严重不均衡问题、异常的脏数据往往在分布上与其它数据差别较大，这些问题往往会导致弱势类别（属性）数据学习不充分、模型歧视、异常数据干扰模型性能等后果。因此，在使用数据开展训练之前，开发者通过对数据进行详细的分布统计，就可以在一定程度上挖掘数据集中存在的不足，进而采取必要的措施缓解相应的问题。

  回归分析是数据分析中常用的一门技术，目的在于了解两个或多个变量间是否相关以及它们之间的相关方向与强度，并建立数学模型以便观察特定变量来预测研究者感兴趣的变量。更具体的来说，回归分析可以帮助人们了解在只有一个自变量变化时因变量的变化量。一般来说，通过回归分析我们可以由给出的自变量估计因变量的条件期望。例如：对原始的数据、数据在模型中的表示（Activation Map）以及模型对于数据的输出进行回归分析，区分出其中正常数据和毒化数据。

  聚类分析是一门在机器学习、数据挖掘等领域经常使用的统计数据分析技术。聚类是把具有相似特征的对象通过分类的方法分为多个子集，这样在同一个子集中的成员对象都有相似的一些属性。聚类是一类无监督的分析方法，在使用聚类的过程中，使用者无需知道数据的标签，而仅仅根据数据具有的特征就可

以初步地对大量数据进行简单分类。数据特征间的相似性往往使用特征之间的距离来度量，例如曼哈顿距离、欧氏距离、马氏距离等。

- **数据预处理**

  在数据筛选和统计分析的基础上，结合现有数据集的特点，根据 AI 模型具体任务，有目的性地对数据集进行相应处理，以满足不同类型的需求。数据预处理包括数据增强、数据规范化、特征变换与去噪等。数据增强是指对当前拥有的数据集进行适当的修改，以提高数据的数量和质量，从而实现模拟真实环境，提高模型泛化能力。它可以对数据较少的部分进行有针对性的扩充，以解决数据量较少、数据分布不均衡等问题。数据的规范化处理是将数据的不同特征进行重新规范，使之落入一个小的特定区间，去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。它可以规范数据量级，消除不同特征指标之间的量纲影响，便于不同单位或量级的指标之间进行比较和加权，加快模型收敛速度，提高模型的精度。特征变换是指通过对图片进行一定的净化处理，减少样本中自然噪音或者恶意修改的影响，例如针对性特征掩盖、图像位深度减少、模糊图像、图像缝合等特征变换技术可以去除冗杂信息，破坏恶意样本中对抗扰动的攻击性能 [95][96]。去噪是对输入样本进行多种清理操作，去除自然环境下产生的噪音，减轻基于扰动的对抗攻击 [98][99][113]。

模型加固　除了数据层面，我们还可以从模型结构层面对 AI 模型进行加固。模型加固是指通过对抗训练、模型压缩、模型蒸馏与后门检测等技术从模型结构、参数层面消除模型漏洞、提升模型安全性，使模型的决策空间更加鲁棒，从而有效防御对抗攻击和后门攻击等。

- **对抗训练** [103][185] 是针对对抗攻击最有效的防御手段之一，也是模型在实际部署时针对对抗攻击首要考虑的防御手段。对抗训练是将 PGD、FGSM 等方法生成的对抗样本作为数据增强策略混入到模型的训练数据集中，通过数据增强使模型学习到对抗样本的特征，从而有效识别出对抗样本。经典的基于 PGD、FGSM 对抗训练模型仍然会受到黑盒攻击，因此可以考虑使用多模型进行集成对抗训练进行防御 [92]。

- **模型压缩**保证模型预测效果的前提下，尽可能地降低模型的大小。模型压缩可以删除模型中一些不必要存在，但可能会被攻击者利用网络的结构和参数。模型压缩主要有防御性蒸馏与模型权重剪枝：防御性蒸馏技术 [101] 不仅可以压缩模型，而且能通过使模型输出更加平滑从而提高模型对于对抗攻击的鲁棒性。防御性蒸馏模型可以有效的防御 FGSM 和 JSMA 对抗等攻击算法；模型权重剪枝也是一个有效的压缩深度学习模型的技术。Ye 等人 [186] 提出了一个

并行的权重剪枝和对抗训练算法，实现了使模型获得对抗攻击鲁棒性的同时对模型进行压缩。

- **后门消除**技术用于检测或删除预训练模型中被攻击者恶意埋入的后门。后门消除可以通过诸如模型蒸馏、剪枝等模型压缩操作，在一定程度上破坏模型中可能嵌入的后门特征，提高模型的鲁棒性；也可以使用 [86][87][88][109] 中提出的后门检测方法，获得后门触发器，然后采取类似对抗训练的方法消除后门。

5.4.2　业务评审

业务评审是指模型在业务系统中上线的评审，这个环节主要回答“业务系统的安全能力是否与这个模型是否匹配”这类问题。业务评审通过流程管理的制度有效地阻拦了安全性弱的模型上线，或者阻拦安全性弱的业务系统使用 AI 模型。

5.4.3　风险治理

风险治理环节属于 AI SDL 的后阶段，一般发生在模型上线后。风险治理从实际的攻击溯源出发，充分看见“威胁”，为上线后的模型定制加固方案，尽最大可能消除潜在的 AI 威胁。

第六章　总结与展望

人工智能技术已广泛应用于生物核身、自动驾驶、语音识别、自然语言处理和博弈等多种场景。人工智能技术在加速传统行业的智能化变革的同时，其安全性问题也越来越被人们关注。聚焦于人工智能安全问题，本白皮书从 AI 模型、AI 数据和 AI 承载系统三个角度系统地总结了人工智能技术所面临的威胁，介绍了面对这些威胁的防御手段，并面向工业界给出了安全的人工智能应用一站式解决方案。

人工智能应用在实际部署时面临对抗攻击、数据投毒攻击和模型窃取攻击等多种潜在威胁。在实际应用场景中，多种 AI 攻击同时存在，我们很难用单一的防御技术来应对现实场景中复杂的威胁。此外，在人工智能的攻防对抗过程中防御是更困难的一方，攻击者可以不断更新攻击技术来突破目前最有效的防御系统，然而新的防御系统却需要考虑现存的所有攻击技术。为了应对实际场景中复杂的威胁以及不断变化的威胁手段，AI 安全研究人员更应从人工智能模型的可解释性等理论角度出发，从根本上解决人工智能模型所面临的安全问题。一方面，研究人员在模型的训练阶段可以通过选取或设计本身具有可解释性的模型，为模型增强泛化能力和鲁棒性；另一方面，研究人员要尝试解释模型的工作原理，即在不改变模型本身的情况下探索模型是如何根据输入样本进行决策的。

人工智能安全技术发展迅猛，各类研究方法与技术路线日新月异，在这种情况下，本书难以覆盖所有学术流派和工业实践，有疏漏之处在所难免，本书只求抛砖引玉，希望各位读者批评指正。

参考文献

[1] Nilesh N. Dalvi, Pedro M. Domingos, Mausam, Sumit K. Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 99–108, 2004.

[2] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733, 2017.

[3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[4] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1625–1634. IEEE Computer Society, 2018.

[5] Tencent Keen Security Lab. Experimental security research of tesla autopilot, 2019. `https://keenlab.tencent.com/en/2019/03/29/Tencent-Keen-Security-Lab-Experimental-Security-Research-of-Tesla-Autopilot/`.

[6] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015*, pages 1322–1333. ACM, 2015.

[7] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *25th USENIX*

Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016, pages 601–618. USENIX Association, 2016.

[8] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017*, pages 506–519. ACM, 2017.

[9] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *Workshop on Artifcial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artifcial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019*, volume 2301 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[10] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses - 21st International Symposium, RAID 2018, Heraklion, Crete, Greece, September 10-12, 2018, Proceedings*, volume 11050 of *Lecture Notes in Computer Science*, pages 273–294. Springer, 2018.

[11] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[12] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society, 2017.

[13] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, pages 2041–2055, 2019.

[14] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition.

In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 1528–1540. ACM, 2016.

[15] Qixue Xiao, Kang Li, Deyue Zhang, and Weilin Xu. Security risks in deep learning implementations. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 123–128. IEEE Computer Society, 2018.

[16] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In Lorenzo Cavallaro, Johannes Kinder, XiaoFeng Wang, and Jonathan Katz, editors, *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, pages 225–240. ACM, 2019.

[17] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artifcial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pages 27–38, 2017.

[18] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 6106–6116, 2018.

[19] Chen Zhu, W. Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 7614–7623, 2019.

[20] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21*, 2018.

[21] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. *arXiv preprint arXiv:1910.00033*, 2019.

[22] Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. *CoRR*, abs/2003.03030, 2020.

[23] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[24] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[25] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[26] Tianhang Zheng, Changyou Chen, and Kui Ren. Distributionally adversarial attack. In *Proceedings of the AAAI Conference on Artifcial Intelligence*, volume 33, pages 2253–2260, 2019.

[27] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pages 372–387. IEEE, 2016.

[28] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artifcial Intelligence, (AAAI-18), the 30th innovative Applications of Artifcial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artifcial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 10–17. AAAI Press, 2018.

[29] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 8322–8333, 2018.

[30] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *CoRR*, abs/1712.09665, 2017.

[31] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face ID system. *CoRR*, abs/1908.08705, 2019.

[32] Javier Galbally, Sébastien Marcel, and Julian Fiérrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014.

[33] Jianwei Yang, Zhen Lei, Dong Yi, and Stan Z. Li. Person-specific face antispoofing with subject domain adaptation. *IEEE Trans. Information Forensics and Security*, 10(4):797–809, 2015.

[34] Seong Joon Oh, Bernt Schiele, and Mario Fritz. Towards reverse-engineering black-box neural networks. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 121–144. Springer, 2019.

[35] Jacson Rodrigues Correia da Silva, Rodrigo Ferreira Berriel, Claudine Badue, Alberto Ferreira de Souza, and Thiago Oliveira-Santos. Copycat CNN: stealing knowledge by persuading confession with random non-labeled data. In *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, pages 1–8. IEEE, 2018.

[36] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 691–706. IEEE, 2019.

[37] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *2019 IEEE Conference on Computer Communications, INFOCOM 2019, Paris, France, April 29 - May 2, 2019*, pages 2512–2520. IEEE, 2019.

[38] Briland Hitaj, Giuseppe Ateniese, and Fernando Pérez-Cruz. Deep models under the GAN: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*, pages 603–618. ACM, 2017.

[39] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 14747–14756, 2019.

[40] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, Udam Saini, Charles A. Sutton, J. Doug Tygar, and Kai Xia. Exploiting machine learning to subvert your spam filter. In *First USENIX Workshop on Large-Scale Exploits and Emergent Threats, LEET '08, San Francisco, CA, USA, April 15, 2008, Proceedings*, 2008.

[41] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[42] Jakub Konecný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *CoRR*, abs/1610.05492, 2016.

[43] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin B. Calo. Analyzing federated learning through an adversarial lens. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 634–643. PMLR, 2019.

[44] Cong Liao, Haoti Zhong, Anna Cinzia Squicciarini, Sencun Zhu, and David J. Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. *CoRR*, abs/1808.10307, 2018.

[45] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[46] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks, 2019.

[47] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *CoRR*, abs/1912.02771, 2019.

[48] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. *CoRR*, abs/1807.00459, 2018.

[49] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2020.

[50] Zhaoyuan Yang, Naresh Iyer, Johan Reimann, and Nurali Virani. Design of intentional backdoors in sequential models. *CoRR*, abs/1902.09972, 2019.

[51] Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pretrained models. In *Annual Conference of the Association for Computational Linguistics (ACL)*, July 2020.

[52] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283. PMLR, 2018.

[53] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 284–293. PMLR, 2018.

[54] Lejla Batina, Shivam Bhasin, Dirmanto Jap, and Stjepan Picek. CSI NN: reverse engineering of neural network architectures through electromagnetic side channel. In *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, pages 515–532. USENIX Association, 2019.

[55] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[56] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pages 15–26. ACM, 2017.

[57] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1378–1387. IEEE Computer Society, 2017.

[58] Chong Xiang, Charles R. Qi, and Bo Li. Generating 3d adversarial point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9136–9144. Computer Vision Foundation / IEEE, 2019.

[59] Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, and Kui Ren. Point-cloud saliency maps. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1598–1606. IEEE, 2019.

[60] Hiromu Yakura and Jun Sakuma. Robust audio adversarial example for a physical attack. In *Proceedings of the Twenty-Eighth International Joint Conference on Artifcial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5334–5341. ijcai.org, 2019.

[61] Nicholas Carlini and David A. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 1–7. IEEE Computer Society, 2018.

[62] Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[63] Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[64] Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In *Machine Learning and Data Mining in Pattern Recognition - 13th International Conference, MLDM 2017, New York, NY, USA,*

July 15-20, 2017, Proceedings, volume 10358 of *Lecture Notes in Computer Science*, pages 262–275. Springer, 2017.

[65] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled. In *Proceedings of the Twenty-Seventh International Joint Conference on Artifcial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4208–4215. ijcai.org, 2018.

[66] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019.

[67] Yao Qin, Nicholas Carlini, Garrison W. Cottrell, Ian J. Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5231–5240. PMLR, 2019.

[68] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, page 125, 2016.

[69] R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 4700–4709, 2018.

[70] Aäron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C. Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. Parallel wavenet: Fast high-fidelity speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3915–3923, 2018.

[71] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 9448–9458, 2019.

[72] General data protection regulation, 2018. `https://gdpr-info.eu/`.

[73] Binghui Wang and Neil Zhenqiang Gong. Stealing hyperparameters in machine learning. In *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*, pages 36–52. IEEE Computer Society, 2018.

[74] Honggang Yu, Kaichen Yang, Teng Zhang, Yun-Yun Tsai, Tsung-Yi Ho, and Yier Jin. Cloudleak: Large-scale deep learning models stealing through adversarial examples. In *Proceedings of Network and Distributed Systems Security Symposium (NDSS)*, 2020.

[75] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N. Asokan. PRADA: protecting against DNN model stealing attacks. In *IEEE European Symposium on Security and Privacy, EuroS&P 2019, Stockholm, Sweden, June 17-19, 2019*, pages 512–527. IEEE, 2019.

[76] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *31st IEEE Computer Security Foundations Symposium, CSF 2018, Oxford, United Kingdom, July 9-12, 2018*, pages 268–282. IEEE Computer Society, 2018.

[77] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019.

[78] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*, pages 634–646. ACM, 2018.

[79] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. Updates-leak: Data set inference and reconstruction attacks in online learning. *arXiv preprint arXiv:1904.01067*, 2019.

[80] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: membership inference attacks against generative models. *PoPETs*, 2019(1):133–152, 2019.

[81] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 739–753. IEEE, 2019.

[82] 360 security. Ai security white paper. Website, 2018. `https://www.freebuf.com/articles/network/162875.html`.

[83] Michael Kissner. Hacking neural networks: A short introduction. *CoRR*, abs/1911.07658, 2019.

[84] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 8011–8021, 2018.

[85] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith Chinthana Ranasinghe, and Surya Nepal. STRIP: a defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference, ACSAC 2019, San Juan, PR, USA, December 09-13, 2019*, pages 113–125. ACM, 2019.

[86] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 707–723. IEEE, 2019.

[87] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial*

Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, pages 4658–4664. ijcai.org, 2019.

[88] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. ABS: scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, pages 1265–1282. ACM, 2019.

[89] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[90] Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[91] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[92] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[93] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *CoRR*, abs/1803.06373, 2018.

[94] Xuanqing Liu and Cho-Jui Hsieh. Rob-gan: Generator, discriminator, and adversarial attacker. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11234–11243. Computer Vision Foundation / IEEE, 2019.

[95] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *6th International*

Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018.

[96] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018.

[97] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[98] Guoqing Jin, Shiwei Shen, Dongming Zhang, Feng Dai, and Yongdong Zhang. APE-GAN: adversarial perturbation elimination with GAN. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 3842–3846. IEEE, 2019.

[99] Dongyu Meng and Hao Chen. Magnet: A two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*, pages 135–147. ACM, 2017.

[100] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1778–1787. IEEE Computer Society, 2018.

[101] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pages 582–597. IEEE Computer Society, 2016.

[102] Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary C. Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[103] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 381–397. Springer, 2018.

[104] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

[105] Chen Wang, Gaoyang Liu, Haojun Huang, Weijie Feng, Kai Peng, and Lizhe Wang. Miasec: Enabling data indistinguishability against membership inference attacks in mlaas. *IEEE Transactions on Sustainable Computing*, 2019.

[106] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, pages 259–274. ACM, 2019.

[107] Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, and Jinwen He. Towards privacy and security of deep learning systems: A survey. 2019.

[108] Manish Kesarwani, Bhaskar Mukhoty, Vijay Arya, and Sameep Mehta. Model extraction warning in mlaas paradigm. In *Proceedings of the 34th Annual Computer Security Applications Conference, ACSAC 2018, San Juan, PR, USA, December 03-07, 2018*, pages 371–380. ACM, 2018.

[109] Ximing Qiao, Yukun Yang, and Hai Li. Defending neural backdoors via generative distribution modeling. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 14004–14013, 2019.

[110] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July*

10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, pages 530–539. PMLR, 2018.

[111] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 8803–8812, 2018.

[112] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[113] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60, 2019.

[114] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[115] Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. Explainable artificial intelligence: A survey. In Karolj Skala, Marko Koricic, Tihana Galinac Grbac, Marina Cicin-Sain, Vlado Sruk, Slobodan Ribaric, Stjepan Gros, Boris Vrdoljak, Mladen Mauher, Edvard Tijan, Predrag Pale, and Matej Janjic, editors, *41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018, Opatija, Croatia, May 21-25, 2018*, pages 210–215. IEEE, 2018.

[116] Zachary C. Lipton. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, 2018.

[117] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

[118] S. Rasoul Safavian and David A. Landgrebe. A survey of decision tree classifier methodology. *IEEE Trans. Systems, Man, and Cybernetics*, 21(3):660–674, 1991.

[119] Jerome H Friedman, Bogdan E Popescu, et al. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

[120] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *CoRR*, abs/1511.01644, 2015.

[121] Jinkyu Kim and John F. Canny. Interpretable learning for self-driving cars by visualizing causal attention. *CoRR*, abs/1703.10631, 2017.

[122] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 39–48. IEEE Computer Society, 2016.

[123] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009.

[124] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*, pages 372–378. IEEE, 2014.

[125] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the frst instructional conference on machine learning*, volume 242, pages 133–142. Piscataway, NJ, 2003.

[126] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[127] Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[128] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[129] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[130] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[131] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1376–1401. JMLR.org, 2015.

[132] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5947–5956, 2017.

[133] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 254–263. PMLR, 2018.

[134] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *CoRR*, abs/1805.12076, 2018.

[135] Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6151–6159, 2017.

[136] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars,

editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer, 2014.

[137] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017.

[138] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banf, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014.

[139] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

[140] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[141] Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579, 2015.

[142] Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[143] Chandan Singh, W. James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[144] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3449–3457. IEEE Computer Society, 2017.

[145] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2950–2958. IEEE, 2019.

[146] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016.

[147] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6967–6976, 2017.

[148] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. Towards interpretable face recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9347–9356. IEEE, 2019.

[149] Guido W Imbens and Donald B Rubin. **人工智能安全**. Cambridge University Press, 2015.

[150] Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi N. R. Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[151] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. *CoRR*, abs/1806.02877, 2018.

[152] Tiago Jose de Carvalho, Fábio Augusto Faria, Hélio Pedrini, Ricardo da Silva Torres, and Anderson Rocha. Illuminant-based transformed spaces for image forensics. *IEEE Trans. Information Forensics and Security*, 11(4):720–733, 2016.

[153] Ali Khodabakhsh, Ramachandra Raghavendra, Kiran B. Raja, Pankaj Shivdayal Wasnik, and Christoph Busch. Fake face detection methods: Can they be generalized? In *2018 International Conference of the Biometrics Special Interest Group, BIOSIG 2018, Darmstadt, Germany, September 26-28, 2018*, volume P-282 of *LNI*, pages 1–6. GI / IEEE, 2018.

[154] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1–11. IEEE, 2019.

[155] Ramachandra Raghavendra, Kiran B. Raja, Sushma Venkatesh, and Christoph Busch. Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1822–1830. IEEE Computer Society, 2017.

[156] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection. *CoRR*, abs/1803.11276, 2018.

[157] David Guera and Edward J. Delp. Deepfake video detection using recurrent neural networks. In *15th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2018, Auckland, New Zealand, November 27-30, 2018*, pages 1–6. IEEE, 2018.

[158] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security, WIFS 2018, Hong Kong, China, December 11-13, 2018*, pages 1–7. IEEE, 2018.

[159] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: automated whitebox testing of deep learning systems. *Commun. ACM*, 62(11):137–145, 2019.

[160] Jianmin Guo, Yu Jiang, Yue Zhao, Quan Chen, and Jiaguang Sun. Dlfuzz: differential fuzzing testing of deep learning systems. In *Proceedings of the 2018 ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/SIGSOFT FSE 2018, Lake Buena Vista, FL, USA, November 04-09, 2018*, pages 739–743. ACM, 2018.

[161] Lei Ma, Felix Juefei-Xu, Fuyuan Zhang, Jiyuan Sun, Minhui Xue tand Bo Li, Chunyang Chen, Ting Su, Li Li, Yang Liu, Jianjun Zhao, and Yadong Wang. Deepgauge: multi-granularity testing criteria for deep learning systems. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, ASE 2018, Montpellier, France, September 3-7, 2018*, pages 120–131. ACM, 2018.

[162] Xiang Ling, Shouling Ji, Jiaxu Zou, Jiannan Wang, Chunming Wu, Bo Li, and Ting Wang. DEEPSEC: A uniform platform for security analysis of deep learning model. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 673–690. IEEE, 2019.

[163] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017.

[164] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018.

[165] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.

[166] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[167] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 10900–10910, 2018.

[168] Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5283–5292. PMLR, 2018.

[169] Eric Wong, Frank R. Schmidt, Jan Hendrik Metzen, and J. Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 8410–8419, 2018.

[170] Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy A. Mann, and Pushmeet Kohli. A dual approach to scalable verification of deep networks. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artifcial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 550–559. AUAI Press, 2018.

[171] Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. Safety verification of deep neural networks. In Rupak Majumdar and Viktor Kuncak, editors, *Computer Aided Verifcation - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I*, volume 10426 of *Lecture Notes in Computer Science*, pages 3–29. Springer, 2017.

[172] Rüdiger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In Deepak D'Souza and K. Narayan Kumar, editors, *Automated Technology for Verifcation and Analysis - 15th International Symposium, ATVA 2017, Pune, India, October 3-6, 2017, Proceedings*, volume 10482 of *Lecture Notes in Computer Science*, pages 269–286. Springer, 2017.

[173] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pages 3578–3586, 2018.

[174] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin T. Vechev. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 10825–10836, 2018.

[175] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy A. Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *CoRR*, abs/1810.12715, 2018.

[176] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Formal security analysis of neural networks using symbolic intervals. In *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, pages 1599–1614. USENIX Association, 2018.

[177] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 656–672. IEEE, 2019.

[178] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 2019.

[179] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[180] Sorelle A. Friedler, Chitradeep Dutta Roy, Carlos Scheidegger, and Dylan Slack. Assessing the local interpretability of machine learning models. *CoRR*, abs/1902.03501, 2019.

[181] Chih-Hong Cheng, Georg Nührenberg, Chung-Hao Huang, Harald Ruess, and Hirotoshi Yasuoka. Towards dependability metrics for neural networks. In *16th*

ACM/IEEE International Conference on Formal Methods and Models for System Design, MEMOCODE 2018, Beijing, China, October 15-18, 2018, pages 43–46. IEEE, 2018.

[182] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2019.

[183] Zhi Quan Zhou, Liqun Sun, Tsong Yueh Chen, and Dave Towey. Metamorphic relations for enhancing system understanding and use. *IEEE Transactions on Software Engineering*, 2018.

[184] Weijie Chen, Berkman Sahiner, Frank Samuelson, Aria Pezeshk, and Nicholas Petrick. Calibration of medical diagnostic classifier scores to the probability of disease. *Statistical methods in medical research*, 27(5):1394–1409, 2018.

[185] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 656–672. IEEE, 2019.

[186] Shaokai Ye, Xue Lin, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, and Yanzhi Wang. Adversarial robustness vs. model compression, or both? In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 111–120. IEEE, 2019.