[第1章 人工智能安全概述 2](#_Toc178280933)

[1.1 人工智能安全的引入 2](#_Toc178280934)

[1.2 人工智能安全的概念 4](#_Toc178280935)

[1.3 人工智能安全的架构、风险及应对方法 5](#_Toc178280936)

[1.3.1 人工智能安全架构 5](#_Toc178280937)

[1.3.2 人工智能安全风险 6](#_Toc178280938)

[1.1.3 人工智能安全风险的应对方法 8](#_Toc178280939)

[1.4 人工智能安全现状 9](#_Toc178280940)

[1.5 教材的组织、学习和讲授方法 10](#_Toc178280941)

[1.5.1 教材的组织 10](#_Toc178280942)

[1.5.2 教材的学习方法 12](#_Toc178280943)

[1.5.3 教材的讲授方法 12](#_Toc178280944)

[1.6 思考题 13](#_Toc178280945)

[参考文献 13](#_Toc178280946)

# 第1章 人工智能安全概述

本章从经典的两个人工智能安全案例事件说起，引入人工智能安全的概念及框架，说明了人工智能安全现状，最后给出了本教材的组织、学习和讲授教材的方法。

* 知识与能力目标

1. 了解人工智能安全的重要性。
2. 认知人工智能安全的概念。
3. 掌握人工智能安全的模型。
4. 了解这门课的知识体系。
5. 了解如何学习这门课程。
6. 熟悉这门课的讲授方法。

## 1.1 人工智能安全的引入

2016年11月18日，在中国深圳举办的“第十八届中国国际高新技术成果交易会”上，一台名为“小胖”的机器人引起了很多人的兴趣。如图1.1所示为小胖机器人。



图1.1 小胖机器人

小胖机器人原为4-12岁儿童研发，主要用于教育目的。然而在展示的过程中这台机器人却突发故障，在没有任何指令的前提下自行打砸展台的玻璃，最终导致部分展台被破坏。更为严重的是，该机器人在破坏展台的过程中还砸伤了一名路人。如图1.2所示为受伤的路人。



图1.2 小胖机器人伤害的路人

该事件大概是国内最早报道的机器人伤害人类的事件。以前类似事件多是在电影电视里才能看到的虚拟场景。从深圳这次安全事件可以看出，人工智能方法自身如果出现问题就可以威胁到人类的安全。

人工智能技术的快速应用，特别是人工智能大模型技术的流行，能促使人工智能技术与实体经济进行加速融合与应用，进而提高人们的生活品质。然而，随着人智能技术的大规模应用，随之而来的安全问题也层出不穷。

当前，人工智能技术应用最具新颖和挑战性的场景之一就是自动驾驶技术，但是就是这样一种技术处理不好的话，依然影响着人类的生命安全。例如，2018年3月23日，苹果公司华裔工程师黄伟伦驾驶特斯拉Model X在加利福尼亚州山景城附近的高速路上，撞上公路屏障不幸身亡。如图1.3所示为事故现场。



图1.3 特斯拉车事故现场

美国国家运输安全委员会（NTSB）的调查结果显示，该车祸发生前，这辆特斯拉车的自动驾驶辅助系统被使用了近19分钟，当时汽车以每小时71英里的速度偏离了高速公路。这场导致驾驶员死亡的车祸明显是技术问题，而非人为因素。从这次事件可以看出，就连特斯拉这样的世界级顶流自动驾驶车企，在面临人工智能安全问题时，依然会出重大安全问题。由此可见，人工智能技术的应用处理不好，会导致人类的生命受到安全威胁。

以上两个典型的与人工智能相关的安全事件，一个是人工智能系统自身的安全问题，也称为人工智能原生安全；另一个是人工智能技术应用出现的安全问题，也称作人工智能衍生安全。以上两种安全问题引起了国内外的广泛讨论，人们纷纷提出一个问题：人工智能是否安全呢？

人工智能安全关系着全人类命运，通过防范安全风险，推进人工智能技术发展与提升人工智能安全治理能力已成为全人类的共识。针对人工智能安全威胁由局部攻击向系统化协同攻击的演化，导致单一的安全检测与防护技术无法应对复合攻击，因此加速提升人工智能安全检测与防护能力，保障人工智能安全刻不容缓。通过在大学开展人工智能安全方面的教育，特别是引导学生加强人工智能安全实践方面的教育具有重要的现实意义。

## 1.2 人工智能安全的概念

人工智能（Artificial Intelligence），英文缩写为AI。它是指通过计算机科学、数学、统计学等多学科交叉融合的方法，开发出模拟人类智能的技术和算法。它通过模拟人类智能的学习、推理、感知和行动能力，实现机器自主的思考和决策，从而完成一系列复杂的任务和功能。人工智能是十分广泛的科学，包括机器人、语言识别、图像识别、自然语言处理、专家系统、机器学习，计算机视觉等。

人工智能安全是指通过釆取必要措施，防范对人工智能系统的攻击、侵入、干扰、破坏和非法使用以及意外事故，使人工智能系统处于稳定可靠运行的状态，以及遵循人工智能以人为本、权责一致等安全原则，保障人工智能算法模型、数据、系统和产品应用的完整性、保密性、可用性、鲁棒性、透明性、公平性和隐私的能力。这个人工智能安全定义是全国信息安全标准化技术委员会在《人工智能安全标准化白皮书(2019年版)》上发布的。以上的对人工智能安全定义只是强调人工智能内在的安全。更广义上的讲，人工智能安全还包括人工智能的应用安全。

## 1.3 人工智能安全的架构、风险及应对方法

本小节从人工智能安全的架构讲起，讲述了人工智能安全主要分为原生安全和衍生安全。然后分析了人工智能安全系统可能存在的安全风险，以及这些安全风险的应对方法。

### 1.3.1 人工智能安全架构

人工智能安全框架如图1.1所示。包含安全目标、安全风险、安全评估和安全保障4大维度。人工智能安全分为4个核心步骤：



图1.1 人工智能安全架构

第1步：从人工智能的应用安全和人工智能系统自身的安全角度来设立安全目标。

第2步：梳理人工智能的原生安全和衍生安全风险。

(1)原生安全(也叫内生安全)主要是人工智能系统内在的安全或本身的安全，包括人工智能系统中机器学习框架的安全、数据安全和算法模型的安全等。

(2)衍生安全主要是人工智能系统应用到实际生活当中而引起的安全问题。

第3步：对人工智能安全现状进行评估。主要是评估数据、算法、基础设施和系统应用所面临的风险程度。

第4步：根据安全评估的结果，综合运用技术和管理相结合的方法保障人工智能系统的安全。

原生安全是人工智能技术自身在可解释性、鲁棒性、可控性、稳定性等方面存在的缺陷。衍生安全是人工智能技术在应用的过程中，由于不当使用或外部攻击造成系统功能失效或错误使用。

### 1.3.2 人工智能安全风险

随着人工智能技术的迅猛发展，越来越多的厂商和组织开始使用人工智能系统，从而为其业务增加智能化、数字化和自动化的元素。虽然人工智能系统的优势非常显著，但人工智能系统的发展也带来了许多安全风险。因此，在使用人工智能系统的过程中，必须注重安全风险的管理，以减少安全风险对企业和组织业务造成的影响。总体而言人工智能安全风险主要包含以下几类：

1.数据安全风险

攻击者利用模型的输出信息可以开展模型盗取攻击和训练数据盗取攻击。在机器学习模型的训练和应用过程中，所使用的训练数据和模型参数都有被泄露的风险。攻击者可以根据要攻击的目标模型，查询样本获取目标攻击模型的预期结果，从而导致模型相关参数泄露，生成替代模型或改进模型，进而构成知识产权方面的侵犯。攻击者也可以推断特定数据集是否用来训练目标模型的，进而获得训练相关信息及训练数据的隐私信息，再使用特定的测试数据进行投毒攻击等。

2.算法模型安全风险

针对人工智能深度学习算法提取样本特征的特点，在不改变深度学习系统模型的前提下，通过构造相关输入样本，使系统输出错误的结果来对抗样本攻击。这种攻击可分为躲避攻击（即非定向攻击）和假冒攻击（即定向攻击）。攻击者可以通过特定样本误导深度学习系统输出特定的错误结果，例如攻击者A可以通过虚假相似样本解锁用户B手机中的人脸识别系统。攻击者也可以改变深度学习系统输出非特定的错误结果，例如攻击者可以控制监控的摄像头，使其实现人员隐身、换人或身份误判等。

3.机器学习框架安全风险

人工智能算法基于机器学习框架完成其模型的搭建、训练和运行。深度学习框架需要依靠于大量的基础库和第三方组件支持，组件的复杂度会严重降低深度学习框架的安全性。例如，由于某个组件开发者的疏忽，或者不同组件开发者之间开发标准的不统一，都可能在深度学习框架中引入漏洞。攻击者可以改写人工智能系统的关键数据，或者通过数据流劫持的方法控制代码执行，实现对人工智能系统的窃取、干扰、控制甚至破坏。

4.人工智能系统应用安全风险。

不当使用、外部攻击和业务设计安全错误等都会引发人工智能系统的应用安全风险。例如，攻击者可以通过提示词技术输入错误数据，使人工智能系统自学习到错误信息；也可以通过智能终端、应用软件和设备的漏洞对人工智能系统实施注入攻击，如以人脸识别系统、智能语音助手、伪造图像为入口攻击后台业务系统；还可以利用应用系统自身设计的缺陷实施攻击，如使用者的操作权限设置不当、应用场景外在风险考虑不够等。

5.法律风险。

人工智能系统在使用过程中可能违反国家的相关法律法规。例如在人脸识别、语音识别、身份认证等领域使用的人工智能技术可能涉及侵犯隐私问题。如果这些个人隐私数据被大量窃取和泄露将会对个人造成严重后果，使用的企业和组织也会被判罚款、整顿或停运等。

### 1.1.3 人工智能安全风险的应对方法

随着人工智能技术和应用的发展，人工智能系统逐渐渗透到人们的生活的保个方面，给我们生活带来了许多便利。在使用人工系统过程中，必须注重风险管理，避免风险对人们生活和工作造成不良影响。加强人工智能系统的风险管理，准备足够的应急响应方法对企业和组织来说非常重要。人工智能系统风险的主要应对方法如下：

1.建立完善的安全管理制度。

三分技术，七分管理，这句话在人工智能安全领域依然适用。在人工智能系统应用过程中，安全管理应该放在首位。因此应建立相对完善的人工智能安全管理制度，制定不同等级的安全应对策略，对每个等级的信息和应用分别使用不同的方法进行保护。

2.人工智能系统风险评估。

对每个人工智能系统应用前都要进行风险评估。评估的目的是确定人工智能系统的安全性和可靠性。通过风险评估，可以有效的识别人工智能系统在应用中面临的各种威胁和隐患，提前采取相应的方法，从而在危险一但发生时，可以有足够的应急措施，进而将损失减少到最低。

3.建立应急响应机制。

在对人工智能系统进行风险评估之后，必须建立相应的风险应对机制。这些机制主要包括：

(1) 预警机制。建立合理的预警机制，提前预测和识别风险，限制风险的发展，尽量减小风险带来的影响。

(2)备份机制。对人工智能系统的重要数据进行备份，防止数据丢失，保证系统在出现问题时，依然能顺利运行。

(3)应急响应机制。不要因为人工智能系统存在安全威胁或可能的隐患而不敢使用它，或者过多地限制使用它。而应该在大胆地使用它的同时，提前给出系统的应急响应方案。一旦人工智能系统突发异常情况，可以做出快速的应对，尽最大努力减少损失。

4.加强法律法规的制定和执行。

在人工智能系统应用的时候，必须加强法律法规的制定和执行。应该借鉴国外人工智能安全法律法规的经验，规定人工智能系统的应用和限制条件，以保障企业和个人数据的安全；同时应该明确责任人违反人工智能安全时应承担的相应处罚。

## 1.4 人工智能安全现状

随着人工智能技术的不断进步，全球人工智能安全治理体系面临着前所未有的挑战！如何构建一个安全、可靠、高效的人工智能安全治理体系，成为各国共同面临的重要课题。人工智能的安全性甚至于关乎整个人类的命运。

人工智能安全与国家安全紧密相连，世界各国都试图在该领域占领先机，美国、欧盟、日本等多个国家接连发布相关政策法规加以规制。2024年3月21日，联合国大会通过了首个关于人工智能安全的全球决议草案，倡议各国合作起来共同开发“安全、可靠和值得信赖的”人工智能系统，为全球各国合作制定并实施人工智能技术和安全应用的标准奠定了坚实基础。2024年3月13日欧盟通过了《人工智能法案》，该法案强调依照不同风险等级来对人工智能系统安全进行管理。2023年10月，美国颁布《关于安全、可靠和值得信赖的人工智能开发和使用的行政命令》，2024年2月美国成立人工智能安全研究所联盟，该联盟得到200多家企业和组织的支持。中国也积极参与并倡导全球人工智能治理工作，中国国家网信办在2023年10月提出《全球人工智能治理倡议》，围绕人工智能发展、安全、治理三方面系统阐述了人工智能治理的中国方案，希望促进人工智能技术造福人类，推动构建人类命运共同体。

目前，我国在人工智能安全领域已取得一系列成果，但人工智能的安全检测、防护、监测预警技术和安全管理仍不够完善，导致人工智能在高安全等级领域应用落地受到一定的制约。人工智能安全治理是一个复杂庞大的系统性工程，需从 人工智能安全理论、标准、检测、防护等多个方面，全面夯实人工智能安全体系。

2023年7月10日,中国国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局公布《生成式人工智能服务管理暂行办法》,自2023年8月15日起施行。该办法旨在促进生成式人工智能健康发展和规范应用，维护国家安全和社会公共利益，保护公民、法人和其他组织的合法权益。

尽管我国在人工智能安全领域虽然已经有一些政策和法规，但是依然不够完善。如何加强我国人工智能安全治理工作，确保人工智能产业与技术的健康发展依然任重道远。

## 1.5 教材的组织、学习和讲授方法

由于关于人工智能安全的知识点非常多，本教材只是从实践应用的角度讲述了一些主要的人工智能安全知识。重点是教会学生如何通过Python语言编程的方法来实现一些经典的人工智能安全案例。

### 1.5.1 教材的组织

教材主要从原理和实践两个方向出发编写知识点，如表1.1所示。这些原理和实践知识都是当前人工智能安全领域的热点。

表1.1 教材的知识点

|  |  |  |
| --- | --- | --- |
| 章数 | 原理知识 | 实践知识 |
| 第1章 | 人工智能安全的概念和模型 |  |
| 第2章 | 生成对抗网络GANs、深度神经网络 | 生成对抗网络模拟Sin曲线、模型窃取 |
| 第3章 | 卷积神经网络 | 数据投毒、人脸活体检测、验证码识别 |
| 第4章 | 对抗样本生成算法 | 图像对抗 |
| 第5章 | 随机森林算法 | 图像去噪 |
| 第6章 | 贝叶斯分类算法、SVM分类算法 | 垃圾邮件过滤 |
| 第7章 | 长短期记忆网络 | 网络攻击检测 |
| 第8章 | 梯度下降算法 | 模型逆向攻击 |
| 第9章 | 深度伪造技术 | 人脸伪造 |
| 第10章 | 影子模型攻击技术 | 成员推理攻击 |
| 第11章 | 神经网络 | 属性推理攻击 |
| 第12章 | 算法歧视 | 模型公平性检测与提升 |
| 第13章 | 深度学习 | 图像水印去除 |
| 第14章 | Tacotron模型、梅尔谱图、长短记忆网络、混合注意力机制 | 语音合成 |
| 第15章 | 深度学习 | 视频检测 |
| 第16章 | 图神经网络、小样本学习、迁移学习、代码属性图 | 代码漏洞挖掘 |

本教材的所有实践内容如表1.2所示。表中给出了所有实践编程的题目，以及他们的难度水平。期中1星级最简单，5星级最难。在做Python实践编程的时候，教师可以根据实践内容的偏好和难度级别选择要做的实践内容。本教材绝大部分编程实践内容都可以在普通笔记本电脑上实现。

表1.2 实践内容列表

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 实践名称 | 难度 | 备注 |
| 1 | 实践2-1基于生成对抗网络的抛物线样本模拟 | 1星级 | 衍生安全 |
| 2 | 实践2-2基于对抗性攻击无数据替代训练的模型窃取 | 5星级 | 原生安全 |
| 3 | 实践3-1基于卷积神经网络的数据投毒 | 3星级 | 原生安全 |
| 4 | 实践3-2基于卷积神经网络的人脸活体检测 | 2星级 | 衍生安全 |
| 5 | 实践3-3 基于卷积神经网络的验证码识别 | 2星级 | 衍生安全 |
| 6 | 实践4-1基于对抗样本生成算法的图像对抗 | 2星级 | 衍生安全 |
| 7 | 实践5-1基于随机森林算法的图像去噪 | 2星级 | 衍生安全 |
| 8 | 实践6-1 基于贝叶斯和SVM分类算法的垃圾邮件过滤 | 1星级 | 衍生安全 |
| 9 | 实践7-1 基于双向LSTM模型的网络攻击检测 | 2星级 | 衍生安全 |
| 10 | 实践8-1基于梯度下降的模型逆向攻击 | 3星级 | 原生安全 |
| 11 | 实践9-1基于深度伪造技术的人脸伪造 | 3星级 | 衍生安全 |
| 12 | 实践10-1基于影子模型的成员推理攻击 | 2星级 | 原生安全 |
| 13 | 实践11-1 基于神经网络的属性推理攻击 | 3星级 | 衍生安全 |
| 14 | 实践12-1 模型公平性检测与提升 | 3星级 | 原生安全 |
| 15 | 实践13-1基于Skip Encoder-Decoder网络的图像水印去除 | 2星级 | 衍生安全 |
| 16 | 实践14-1基于Tacotron2的语音合成 | 3星级 | 衍生安全 |
| 17 | 实践15-1基于YOLOv5的安全帽识别实验 | 4星级 | 衍生安全 |
| 18 | 实践16-1基于图神经网络的代码漏洞检测 | 5星级 | 衍生安全 |

本教材中所有18个Python编程实践的内容都提供源代码(见本教材的网盘)、数据集和相关库文件。大部分源代码都可以通过复制粘贴的方式进行编程，核心代码采用图片的方式呈现，需要学生自己理解并输入。

本教材的第2个实践内容“实践2-2基于对抗性攻击无数据替代训练的模型窃取”在运行时，需要大量的运算资源进行训练，它的难度是5星级。感兴趣的同学可以在老师的带领下在网上购买计算资源再训练得出实践结果，或者直接使用本书已经训练好的模型(详见本书的网盘)进行实践，得出实践结果。

本教材的第18个实践内容“实践16-1基于图神经网络的代码漏洞检测”需要在Ubuntu虚拟机上运行，并且需要学习的知识点比较多，它的难度是5星级。学生在做实践的时候，如果感觉比较困难，可以以不多于4人的团队形式进行编程实践。

### 1.5.2 教材的学习方法

学生在学习本教材的实践内容时，可以通过以下步骤进行：

第1步：学习教材中实践内容的原理，理解实践的目的。

第2步：下载本教材指定的网盘中的资料，特别是实践内容的原代码、库文件和数据集。

第3步：配置编程实践环境。

第4步：按照教材中指定的实践步骤进行编程实践，直到最终出现正确的实践结果。

第5步：对实践内容进行扩展练习，在教材内容基础上进行更为深入的编程实践。注意：这一步不是必须的，但是如果有的话，会加分。

第6步：按照老师给的实践报告模板撰写实践报告。报告要求如下：

(1)对实践中的原理部分进行更为详细的描述；

(2)报告中需要详细说明自己的实践步骤；

(3)详细说明每行代码的功能作用。

第7步：给老师提交报告。

### 1.5.3 教材的讲授方法

老师在辅导学生做编程实践的时候，可以通过以下步骤进行：

第1步：辅导学生在本教材指定的网盘下载实践所需要的素材，如实践所需要的数据集、图像、视频、模型等。

第2步：辅导学生安装实践用的实践环境，主要是Python编程环境和相关的库文件。

第3步：让学生自己用Python编程实践。大部分源代码都可以从下载的文件中获取并直接复制粘贴，少量核心代码以图片的形式出现在下载的文件当中，需要学生自己理解并输入。

第4步：老师给出学生的实践成绩。老师给出成绩的标准建议如下：

(1)编程实践结果是否正确。成绩在0至60分之间。

(2)查看学生对Python编程实践每行代码的解释程度。成绩在0分至20分之间。

(3)查看学生对于实践内容是否有扩展。例如学生是否在实践报告里对于实践内容的原理部分有更多的篇幅介绍；学生是否在课程实践内容的基础上还有其它相关扩展实践；学生是否在实践报告后面有自己的心得体会等。此部分成绩在0分至20分之间。

(4)以上三部分加起来是100分。

最后强调本教材的内容仅用于教育和学术交流目的，如果用于非法目的，后果自负。

## 1.6 思考题

1.什么是人工智能安全？

2.什么是人工智能的原生安全？

3.什么是人工智能的衍生安全？

## 参考文献

[1] 方滨兴等，人工智能安全[M]. 北京：电子工业出版社,2020.

[2] 李进，谭毓安，人工智能安全基础[M]. 北京：机械工业出版社,2024.

[3] 曾剑平，人工智能安全[M]. 北京：清华大学出版社2022.