第一章 人工智能安全概述

1.1 人工智能简介

智能（Intelligence）指的是具备推理、理解、计划、解决问题、抽象思维、表达意念以及语言和学习的能力，主要与大脑的思维活动有关[1]。人类大脑作为智能的载体，在长期进化过程中形成了高度复杂且精密的结构。现代科学的发展，让我们对大脑的智能有了深入的了解。

人工智能（Artificial Intelligence，简称 AI）是一种模拟人类智能的技术，通过建立可模拟和扩展人类智能的系统，使机器能够具备学习、决策等类人的智能本领[2]。人工智能是一个宽泛的研究领域，涵盖了机器学习、专家系统、深度学习、强化学习、计算机视觉、自然语言处理等多种研究方向。人工智能的核心任务包括感知、学习、推理和应用，从数据中提取模式和规律，进行自主学习和优化，实现自动决策和智能应用。

人工智能在现实生活中的应用已经非常广泛，正在改变各行各业的运作方式。在医疗领域，人工智能正在医疗诊断、药物研发、个性化治疗等方面发挥作用；在金融领域，人工智能可用于风险评估、交易预测、智能顾投、客户服务等；在交通领域，人工智能正用于智能驾驶、交通监控、车辆识别等方面；在教育领域，人工智能技术（如自然语言处理）实现了智能辅导和个性化教育；此外，人工智能还在生活、娱乐、农业、能源等行业发挥重要作用，为人类提供了更高效、智能和便利的服务。

人工智能的发展历程可以追溯到上世纪 50 年代，当时的科学家们研究如何使计算机具备类似于人类智能的能力。随着计算机硬件和算法的发展，人工智能逐渐取得了一些突破性进展，如图 1-1 所示，人工智能研究领域经历了几个关键阶段。

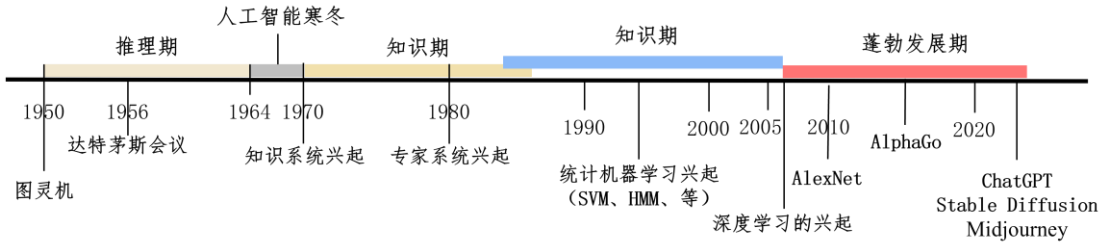

图 1-1 人工智能发展历程

人工智能的概念早在 20 世纪 40 年代就开始萌芽。计算机科学的奠基人阿

兰·图灵在 1950 年提出了著名的“图灵测试”[3]，该测试旨在评估机器是否具有智能。图灵的工作激发了对智能机器的广泛兴趣，并为后来的研究奠定了基础。

在 1956 年的达特茅斯会议上，约翰·麦卡锡首次提出“人工智能”这一术语，标志着人工智能学科的正式诞生。早期的人工智能研究集中在问题求解和逻辑推理等领域，产生了一些重要的系统和程序，如“逻辑理论家”[4]等智能系统。

然而，早期的人工智能领域也经历了挫折，在 20 世纪 70 年代，由于技术和计算资源的限制，以及人们过高的预期未能够实现，人工智能研究进展放缓，这段时期被称为“人工智能寒冬”。尽管如此，全球仍然出现了一些重要的理论与方法，如神经网络的早期探索[5]和专家系统[6]的研发。

20 世纪 80 年代，人工智能研究的重心转向了知识工程和专家系统。这些系统通过将专家知识编码到计算机程序中，成功地解决了一些特定领域的问题，如医疗诊断和地质勘探等。然而，这些系统在处理动态和不确定环境方面仍存在较大的局限性。

进入 90 年代，人工智能领域迎来了机器学习和统计方法的复兴。随着计算能力的提高和大数据的出现，研究者们开发了许多新的算法，如支持向量机（SVM）[7]、隐马尔可夫模型（HMM）[8]、贝叶斯网络[9]等，这些技术被广泛应用于自然语言处理、计算机视觉和语音识别等领域。

2010 年代，基于多层神经网络的深度学习引发了一场革命。深度学习在图像识别、语音识别、自然语言处理等领域取得了突破性进展。AlexNet[10]在 ImageNet 图像分类竞赛中的胜利，标志着深度学习的崛起。谷歌的 AlphaGo 在围棋比赛中击败人类顶级棋手，再次展示了深度学习的强大能力。

进入 21 世纪的第二个十年，随着网络计算能力、互联网和大数据等技术的快速发展，人工智能进入了加速发展的新阶段。基于生成式预训练的 ChatGPT[11]、AI 绘画工具 Stable Diffusion[12]和 Midjourney[13]等大模型，改变了人们的固有认知，进一步解放了人类的脑力，推动了生产力的发展。

从早期的理论探索到当前的大规模应用，人工智能不断改变了人类的生活方式，在自动驾驶、智能医疗、智能客服和金融科技等领域，人工智能技术在不断提高人们的生产效率和创造能力，它已经成为科技创新的重要驱动力。

1.2 人工智能安全

人工智能在飞速发展的同时，也带来了诸多挑战，在可解释性、透明性、公平性等方面，人工智能还存在很大的安全隐患。尤其是“算法黑箱”问题，使得人工智能决策难以被理解和信任。此外，人工智能技术对海量数据的依赖，带来了数据安全、隐私保护和系统漏洞等风险，这些问题可能对国家安全、社会稳定和个人财产构成严重威胁。研究者们正在积极探索新技术，以避免人工智能带来的安全风险。

人工智能安全是指通过采取必要措施，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预防意外事故，从而使人工智能系统处于稳定可靠运行的状态。同时，人工智能安全还要求以人为本、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[14]。

本书主要关注人工智能数据与模型等方面的安全技术，如图 1-2 所示，涉及的具体技术包括：对抗样本[15]、数据投毒[16]、后门攻防[17]、模型水印[18]、数据窃取[19]、深度伪造[20]等。

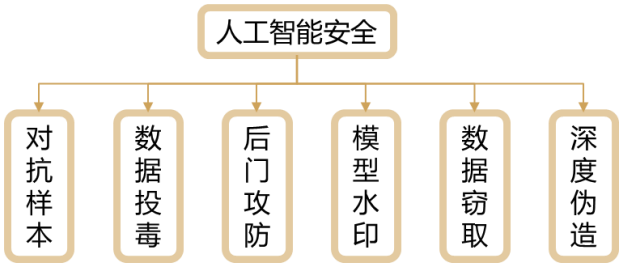

图 1-2 人工智能发展历程

对抗样本（Adversarial Example）是指攻击者可以通过对输入数据进行细微且特定的修改，使得人工智能模型输出错误结果的技术。这些修改通常对于人类来说是难以察觉的，但足以使模型做出错误的分类或预测。对抗样本对图像分类、语音识别等领域的模型构成了严重威胁。例如，在图像识别中，通过在一张猫的图片上添加微小的噪声，可以使模型将其错误识别为狗。这种攻击方式不仅在实验室环境中有效，在现实世界中也有潜在的危险，例如通过修改交通标志的图案来误导自动驾驶系统等。

数据投毒（Data Poisoning）与对抗样本不同，数据投毒着眼于通过操控训练数据集来影响模型的表现。攻击者对数据集进行修改，当模型训练者使用中毒数

据集训练模型后，得到的模型会对特定的输入产生误判。这种攻击通过操纵数据达到操纵模型的目的，可能给模型的性能带来很大的不确定性。

后门攻防（Backdoor）是指攻击者在模型训练阶段插入恶意代码或数据，使得模型在特定触发条件下输出攻击者预期的结果。这种攻击在模型部署后难以被察觉，但可能带来严重后果。例如攻击者在训练自动驾驶模型时，插入了带有特定标志的训练数据，使得车辆在识别到该标志时做出错误的驾驶决策，例如突然加速或转向，导致严重的安全隐患。

模型水印（Model Watermarking）用于保护人工智能模型的知识产权。通过在模型中嵌入特定的标识，可以验证模型的归属，防止未经授权的复制和使用。例如在一个图像分类模型中嵌入独特的数字水印，这些水印在模型的输出中表现出来，能够证明模型的归属权。假如某公司发现其模型被盗用，可以通过验证输出中的水印来确认侵权行为。

数据窃取（Data Theft）是指攻击者通过分析人工智能系统的中间环节信息来窃取模型训练中使用的数据，或通过查询已部署的模型来推断原始训练数据的内容。例如，在医疗场景中，通过分析人工智能模型，攻击者可能推断出患者的病情和隐私信息，造成严重的安全和隐私威胁。

深度伪造攻防（Deepfake）指的是利用深度学习技术生成逼真的假视频和音频，可能被用于欺诈、诽谤等不法行为。深度伪造对个人隐私、公共安全和信息可信度构成了严重威胁。例如，通过深度伪造技术生成政治人物的假视频，传播虚假信息以影响公众舆论和选举结果。深度伪造检测技术致力于识别和揭露这些虚假内容，确保信息的真实性和可信度。

此外，随着人工智能技术渗透到医疗、交通、军事、社会服务等领域，与此同时也会引发一系列伦理道德问题。例如，在医疗领域，尽管人工智能提高了诊断和治疗效率，但也引发了关于医疗决策权的伦理争议，尤其是在人工智能诊断错误时，责任归属将变得模糊。在交通领域，自动驾驶汽车虽然提高了行驶效率，但事故发生时，如何界定责任也将成为争议焦点。同样，在军事领域，自主武器系统的应用引发了人权和国际法的广泛讨论，这类武器如果造成无辜伤亡，责任追究极为复杂。这些问题都显示了人工智能技术在伦理道德方面的脆弱性和潜在风险。面对这些挑战，我们必须确保人工智能技术的使用符合道德和法律准则。

这不仅需要技术上的进步，更需要在法律、伦理和社会层面进行广泛讨论和长期评估，以确保人工智能发展与人类福祉的平衡。

1.3 本章小结

随着人工智能的不断发展，如何保障系统和应用的安全性是全世界关注的焦点。开展人工智能安全研究，将有助于提升人工智能系统的可靠性和信任度。人工智能安全也是网络空间安全与国家安全研究中的一个重要领域，涉及多个方面的知识和方法。本书重点介绍对抗样本、数据投毒、后门攻防、模型水印、数据窃取、深度伪造等方面的人工智能安全攻防技术，这也是当前研究较为广泛的几个方向，让大家更好地认识人工智能安全，引发对人工智能技术本身的思考，进而更好地推动人工智能技术的发展和应用。

课后习题

1. 通过查阅资料，了解人工智能的具体发展历程。
2. 列举几种你身边可能出现的人工智能安全风险隐患。

参考文献

[1] https://en.wikipedia.org/wiki/Intelligence.

[2] https://en.wikipedia.org/wiki/Artificial_intelligence.

[3] Turing A M. Computing machinery and intelligence[M]. Springer Netherlands, 2009.

[4] https://en.wikipedia.org/wiki/Logic_Theorist.

[5] Hayes-Roth F. The knowledge-based expert system: A tutorial[J]. Computer, 1984, 17(09): 11-28.

[6] Werbos P J. Backpropagation through time: what it does and how to do it[J]. Proceedings of the IEEE, 1990, 78(10): 1550-1560.

[7] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[8] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[9] Pearl J. Bayesian networks: A model of self-activated memory for evidential reasoning[C]//Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA, USA, 1985: 329-334.

[10] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems, 2012, 25: 1097-1105.

[11] https://openai.com/.

[12] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10684-10695.

[13] https://www.midjourny.cn.

[14] 中国电子技术标准化研究院. 人工智能安全标准化白皮书（2019 版）[R]. 北京: 中国电子技术标准化研究院, 2019. https://www.tc260.org.cn/file/rgznaqbz.pdf.

[15] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks[C]//2nd International Conference on Learning Representations, ICLR 2014.

[16] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines[C]//Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Madison, WI, USA, 2012: 1467-1474.

[17] Gu T, Dolan-Gavitt B, Garg S. Identifying vulnerabilities in the machine learning model supply chain[C]//Proceedings of the Neural Information Processing Symposium Workshop on Machine Learning and Security, MLSec 2017.

[18] Uchida Y, Nagai Y, Sakazawa S, Satoh S. Embedding watermarks into deep neural networks[C]//Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017: 269-277.

[19] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS 2015: 1322-1333.

[20] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]//Proceedings of the International Conference on Learning

Representations, ICLR 2015.