

面向计算机专业的可信人工智能教学设计

王亚

湖南工业大学, 湖南 株洲 412007

[摘要]我国当前正处于人工智能技术快速发展的关键时期,这一技术正在深刻地改变着我们的生活和工作方式。作为计算机专业的学生,我们需要深入了解我国当前人工智能发展的现状和面临的挑战。我国人工智能技术已经取得了一系列重大进展,涵盖了从理论研究到技术应用的各个方面。然而,我国人工智能发展也面临着一系列挑战,如人才短缺、技术标准不足、数据安全和隐私保护等问题。同时,人工智能技术的不断普及和应用也带来了一系列伦理、法律和社会问题,亟待解决和规范。因此,我们需要深入了解人工智能技术的发展趋势和面临的挑战,积极参与到人工智能技术的研究和应用中,推动人工智能技术的可持续发展,为构建数字化、智能化的社会作出贡献。

[关键词]计算机专业;可信人工智能;教学设计

DOI: 10.33142/fme.v5i3.13080

中图分类号: G642

文献标识码: A

Trusted Artificial Intelligence Teaching Design for Computer Major

WANG Ya

Hunan University of Technology, Zhuzhou, Hunan, 412007, China

Abstract: China is currently in a critical period of rapid development of artificial intelligence technology, which is profoundly changing our way of life and work. As students majoring in computer science, we need to have a deep understanding of the current status and challenges of artificial intelligence development in China. Chinese artificial intelligence technology has made a series of significant progress, covering various aspects from theoretical research to technological applications. However, the development of artificial intelligence in China also faces a series of challenges, such as talent shortage, insufficient technical standards, data security and privacy protection. Meanwhile, the continuous popularization and application of artificial intelligence technology have also brought about a series of ethical, legal, and social issues that urgently need to be addressed and regulated. Therefore, we need to have a deep understanding of the development trends and challenges faced by artificial intelligence technology, actively participate in the research and application of artificial intelligence technology, promote the sustainable development of artificial intelligence technology, and contribute to building a digital and intelligent society.

Keywords: computer major; trusted artificial intelligence; instructional design

引言

人工智能技术正日益渗透到我们生活的方方面面,从语音助手到自动驾驶汽车,都展示了其无可置疑的潜力和影响力。然而,随着人工智能技术的不断发展,我们也必须认识到其中存在的伦理、法律、安全等一系列问题。特别是在计算机专业领域,了解并解决这些问题至关重要。本课程将致力于为计算机专业的学生提供一种全面了解可信人工智能的教学体验。我们将深入探讨人工智能技术背后的伦理和法律考量,探讨如何提高模型的透明性和可解释性,以及如何确保人工智能系统的公平性和消除偏见。我们还将讨论如何增强人工智能系统的安全性和鲁棒性,以抵御各种潜在的攻击和威胁。通过本课程的学习,学生将不仅掌握人工智能技术的核心概念和技术,还将培养批判性思维和团队合作精神,以更好地应对未来的挑战和机遇。我们相信,通过全面了解可信人工智能的教育,我们将为建设一个更加安全、公平和可信的人工智能时代作出贡献。

1 可信人工智能的重要性

可信人工智能的重要性体现在其对社会、经济和技术发展的全方位影响,随着人工智能技术在医疗、金融、自动驾驶等关键领域的迅速扩展,确保这些系统的可信性已成为一个至关重要的问题。这不仅仅涉及到系统的性能和准确性,还关系到公众对人工智能技术的信任和依赖。可信人工智能要求系统在设计 and 运行过程中具备透明性、公平性、可解释性和安全性,能够在多样化和复杂的现实环境中保持稳定和可靠。通过实现透明性,用户和开发者能够理解和监督人工智能决策过程,从而建立信任;通过实现公平性,系统可以避免在种族、性别等方面的偏见,确保决策的公正性;通过实现可解释性,用户可以理解人工智能的决策理由,从而提升系统的可接受性和可控性;通过实现安全性,系统能够抵御各种对抗性攻击和数据泄露,保护用户隐私和数据安全。这些特性不仅可以防止潜在的伦理和法律问题,还能够增强公众和各行业对人工智能技术的信任,促进技术的广泛接受和应用,推动技术创新和

社会进步。

2 面向计算机专业的可信人工智能教学设计

2.1 伦理与法律

2.1.1 人工智能伦理

人工智能伦理是指在人工智能技术发展和应用过程中涉及的伦理原则、价值观和道德规范。随着人工智能技术的快速发展和普及，伦理问题日益受到关注。人工智能伦理旨在解决由人工智能技术可能引发的一系列伦理困境和道德挑战，以确保人工智能技术的发展和符合人类价值观和道德标准，促进技术的良性发展和社会的可持续进步。在人工智能伦理中，涉及到多个关键问题。首先是权责问题，即人工智能系统如何分配权利和责任。在自动驾驶汽车等领域，如果发生事故，责任应归属于谁？这涉及到法律、伦理和社会政策等多方面的考量。其次是公平和正义问题，即人工智能系统的设计和应用是否具有公平性和正义性。例如，人工智能算法在招聘和候选人筛选中是否存在性别、种族等偏见？另外还有隐私和透明度问题，即人工智能系统对用户数据的使用是否符合隐私保护原则，人工智能决策的逻辑是否可以被解释和理解等^[1]。解决人工智能伦理问题需要跨学科的合作，涉及到哲学、社会科学、法律等多个领域。同时，需要制定相关的法律法规和伦理准则，建立健全的监管机制，加强技术研发和应用中的道德教育和风险评估，引导人工智能技术的良性发展，推动社会朝着更加智能、公平和可持续的方向发展。

2.1.2 数据隐私

数据隐私是在人工智能应用中日益突出的一个重要问题，涉及到个人数据的收集、存储、处理和共享等方面。随着信息技术的迅速发展，大量的个人数据被数字化并广泛应用于各个领域，从而引发了对数据隐私保护的关注。首先，数据隐私的重要性体现在个人权利和自由的保护上。个人数据包含了个体的身份信息、偏好、行为习惯等敏感信息，保护这些数据的隐私是维护个人权利和尊严的重要保障。如果个人数据被滥用或泄露，可能导致个人隐私权受到侵犯，甚至引发个人信息泄露、身份盗窃等严重后果。其次，数据隐私的保护也关乎社会稳定和经济发展。在数字化时代，数据已经成为信息社会的核心资源，对数据的合理利用和保护关系到社会和经济的可持续发展。如果数据隐私得不到有效保护，可能导致个人对科技和商业应用的信任降低，阻碍人工智能技术的发展和应用，影响社会经济的健康发展。因此，保护数据隐私是人工智能发展中的一项重要任务。首先需要加强相关的法律法规和标准制定，建立起完善的数据隐私保护体系和监管机制。其次，需要加强数据安全技术的研究和应用，采取有效的加密、匿名化等技术手段，确保个人数据在采集、传输和存储过程中的安全性。同时，也需要加强个人数据主体的权利意识和教育，提高个人数据保护意识，积极维护自身的数据

隐私权益。

2.1.3 法律法规

在计算机专业的可信人工智能教学设计中，法律法规的重要性不可低估。随着人工智能技术的迅速发展和应用，各国纷纷制定了一系列法律法规来规范人工智能技术的发展和应用，以保障公众利益、个人隐私和社会稳定。首先，法律法规为人工智能的发展提供了指导和约束。其次，法律法规为人工智能技术的应用提供了法律责任和监管机制。例如，在自动驾驶汽车等领域，各国政府通过立法明确了自动驾驶汽车的法律责任和标准，规范了自动驾驶汽车的研发和应用。在医疗人工智能等领域，也需要遵循一系列法律法规来保障医疗数据的安全和隐私，确保医疗人工智能系统的准确性和可靠性。此外，法律法规还为人工智能技术的监管和治理提供了法律依据。例如，一些国家和地区通过设立数据保护局、人工智能伦理委员会等机构来监督和管理人工智能技术的发展和应用，加强了对人工智能技术的监管和治理。

2.2 模型透明性与可解释性

2.2.1 模型透明性

在计算机专业的可信人工智能教学设计中，模型透明性是一个关键的概念。模型透明性指的是人们能够理解和解释人工智能模型的工作原理、决策过程和结果。在许多应用中，特别是对于涉及到个人权益和社会公平的领域，模型透明性至关重要。首先，模型透明性有助于提高人们对人工智能技术的信任。当人们能够理解人工智能模型是如何做出决策的，他们会更有信心相信这些决策的准确性和合理性。例如，在医疗诊断中，医生需要知道人工智能模型是如何分析患者的医疗影像并给出诊断结果的，以便更好地理解信任这些结果。其次，模型透明性有助于发现和解决模型中的偏见和错误。当人们能够审查和分析模型的工作原理和决策过程时，他们可以更容易地发现模型中可能存在的偏见和错误。例如，在贷款审批系统中，如果模型基于种族或性别等因素做出不公平的决策，那么这些偏见会在模型透明性的审查过程中被发现并加以修正^[2]。此外，模型透明性还有助于教育和培训人们理解人工智能技术的工作原理和应用场景。通过了解人工智能模型是如何从数据中学习和做出决策的，人们可以更好地应用和创新人工智能技术，推动技术的进步和发展。

2.2.2 可解释性技术

可解释性技术在人工智能领域中扮演着至关重要的角色，它们旨在使人们能够理解和解释机器学习模型的决策过程和结果。在计算机专业的可信人工智能教学设计中，强调可解释性技术的重要性，有助于培养学生的批判性思维和对人工智能系统的信任。可解释性技术的一个重要方向是提供模型的局部和全局解释。局部解释帮助理解单个预测或决策的原因，而全局解释则提供对整个模型行为的

总体理解。例如，对于图像分类模型，局部解释技术可以帮助识别出特定像素对于分类结果的贡献程度，而全局解释技术则可以揭示模型对不同特征的整体依赖性。另一个重要的可解释性技术是基于规则的解释方法，如决策树和规则集。这些方法生成易于理解的规则集，可以直接解释模型的决策逻辑。例如，决策树模型可以通过一系列简单的 if-then 规则来解释每个决策的基础，使用户能够理解模型是如何根据输入特征进行分类或预测的。同时，近年来深度学习模型的解释性也受到了越来越多的关注。虽然深度学习模型通常被认为是黑盒模型，但研究人员已经提出了一系列方法来解释这些模型的决策过程，如梯度相关的方法、激活最大化和特征可视化等。这些方法虽然仍在不断发展，但已经为理解深度学习模型的内部工作原理提供了一定的启示。

2.3 公平性与偏见

2.3.1 公平性

在计算机专业的可信人工智能教学设计中，公平性是一个至关重要的议题。公平性意味着人工智能系统在决策和应用过程中不应受到个体的人种、性别、年龄、社会地位等特征的影响，决策应当基于客观的、公正的标准，而不是歧视性的偏见或刻板印象。确保人工智能系统的公平性需要从多个方面进行考量和实践。首先是数据收集和处理阶段。数据在训练机器学习模型时起着至关重要的作用，因此，必须确保数据集具有代表性和多样性，不偏向任何特定群体。此外，应该通过数据预处理技术来识别和纠正数据中的潜在偏见，以减少偏见在模型中的传递。其次是模型设计和评估阶段。在构建人工智能模型时，应该考虑到公平性原则，并采取相应的措施来避免或减轻模型中的偏见。例如，可以使用公平性约束或惩罚来约束模型对于不同群体的预测结果。同时，模型的公平性应该作为评估指标之一，并且需要不断地对模型进行监测和调整，以确保模型在实际应用中的公平性和稳健性。

2.3.2 消除偏见的方法

在计算机专业的可信人工智能教学设计中，消除偏见是保障人工智能系统公平性的关键一环。偏见可能源自于数据、算法或者应用环境，因此需要采取多种方法来减少或消除偏见。首先，数据的收集和处理阶段是消除偏见的核心。在构建人工智能模型时，应确保训练数据集具有多样性和代表性，涵盖各种不同的群体和场景。此外，需要对数据进行仔细的清洗和预处理，以识别和纠正其中的偏见和不平衡。例如，可以通过数据重采样、增强少数群体样本，或者引入权重等方式来处理数据中的偏见。其次，算法设计和评估阶段也是消除偏见的重要环节。在设计人工智能算法时，应该考虑到潜在的偏见，并采取相应的措施来减轻或避免偏见的影响。例如，可以使用公平性约束或者惩罚来调整模型的预测结果，使其对不同群体更加公

平^[3]。此外，还可以采用多视角的评估方法，从不同群体的角度来评估模型的公平性。最后，社会和政策层面的干预也是消除偏见的重要手段。政府、学术界和行业组织可以制定相关的法律法规和行业标准，要求人工智能系统在设计和应用过程中遵循公平原则，并对违反公平原则的行为进行惩罚或监督。同时，社会各界也可以通过教育、宣传和舆论监督等方式来促进人工智能技术的公平应用，共同建立一个公平、公正的人工智能生态系统。

2.4 安全性与鲁棒性

2.4.1 模型鲁棒性

在计算机专业的可信人工智能教学设计中，模型鲁棒性是确保人工智能系统在面对不同环境和攻击时保持稳定性和可靠性的重要概念。模型鲁棒性指的是模型对于输入数据的变化或扰动的抵抗能力，以及对于对抗性攻击和恶意操纵的防御能力。模型鲁棒性的重要性体现在多个方面。首先，模型鲁棒性能够保护人工智能系统免受意外输入数据的影响。在实际应用中，输入数据可能存在噪声、错误或异常情况，模型鲁棒性可以保证系统能够稳定地处理这些情况，不会因为输入数据的变化而产生不确定或错误的结果。其次，模型鲁棒性能够提高系统对于对抗性攻击的抵抗能力。对抗性攻击是指恶意攻击者通过精心设计的输入数据来欺骗模型，使其产生错误的预测或决策。模型鲁棒性可以通过设计抗攻击的模型结构、采用对抗性训练技术，或者引入对抗性样本检测方法等方式来防御对抗性攻击，保护人工智能系统的安全性和可靠性。最后，模型鲁棒性还能够提高系统的适应能力和泛化能力。在面对未知的环境或数据时，模型鲁棒性可以保证系统能够有效地适应和学习，不会因为环境的变化而失效或退化。这对于人工智能系统在实际应用中的长期稳定性和可持续性具有重要意义。

2.4.2 安全防护

在计算机专业的可信人工智能教学设计中，安全防护是确保人工智能系统在设计、开发和应用过程中免受恶意攻击和数据泄露等安全威胁的关键环节。安全防护涵盖了多个方面，包括数据安全、系统安全和通信安全等。首先，数据安全是保护人工智能系统的核心。数据是人工智能系统的基础，因此必须确保数据在采集、存储、传输和处理过程中的安全性。这包括对数据进行加密、匿名化和访问控制，以防止未经授权的访问和篡改。此外，还需要建立完善的数据安全管理机制，对数据的收集和使用进行严格的监督和管控。其次，系统安全是保护人工智能系统的重要一环^[4]。系统安全涉及到人工智能系统的软硬件安全、身份认证、访问控制等方面。必须确保系统的软件和硬件组件都具有较高的安全性和可靠性，防止恶意软件和恶意攻击者对系统进行入侵和破坏。同时，还需要建立健全的身份认证和访问控制机制，对系统的使用者进行身份验证

和授权,防止未经授权的用户访问系统。最后,通信安全是保护人工智能系统的重要环节。通信安全涉及到系统之间或系统与用户之间的数据传输过程中的安全保障。必须采取加密、数字签名、数据完整性验证等技术手段来保护数据在传输过程中的安全性,防止数据被窃取、篡改或截获。

3 结语

通过本课程的学习,我们深入了解了可信人工智能的重要性以及构建可信人工智能系统所面临的挑战和解决方案。我们探讨了伦理与法律、模型透明性与可解释性、公平性与偏见、安全性与鲁棒性等关键议题,从多个角度审视了人工智能技术的发展现状和未来趋势。在未来的人工智能领域,我们将面临着更多的挑战和机遇。我们需要不断学习和掌握新的知识和技术,不断提升自己的专业能力和素养,以适应不断变化的技术和市场需求。我们还需要积极参与到人工智能技术的研究和应用中,发挥自己的创造力和想象力,推动人工智能技术的创新和进步。最后,让我们携手共建一个安全、公平、可信的人工智能时代,

为构建数字化、智能化的社会做出我们的贡献。愿我们的努力和付出能够为未来的人工智能技术发展和应用带来更多的希望和可能性。

[参考文献]

- [1]陈川,郑子彬,杨晶华,等.面向计算机专业的可信人工智能教学设计[J].计算机教育,2024,12(4):105-108.
- [2]王素月.人工智能教育应用背景下教师教学伦理研究[D].重庆:西南大学,2021.
- [3]张新,于重重,李悦,等.人工智能课程实践教学案例设计——以离散数学课程为例[J].电脑与信息技术,2024,32(2):18-22.
- [4]刘林东,邬依林,张谦.人工智能技术应用背景下程序设计基础课程教学改革探索[J].电脑知识与技术,2024,20(7):143-145.

作者简介:王亚(1982.12—),所学专业:计算机科学与技术,当前就职单位:湖南工业大学,职务:教师,职称级别:讲师。