人工智能

语文出版社

课时分配表

|  |  |  |  |
| --- | --- | --- | --- |
| **章序** | **课程内容** | **课时** | **备注** |
| **1** | 初识人工智能 | 6 |  |
| **2** | 人工智能应用 | 7 |  |
| **3** | 新一代人工智能生态 | 6 |  |
| **4** | 大模型实践 | 6 |  |
| **5** | 人工智能伦理 | 8 |  |
| **6** | 机器学习 | 3 |  |
| **总计** |  | 36 |  |

# 第5课 人工智能伦理

|  |  |  |
| --- | --- | --- |
| **课 题** | 人工智能伦理 | |
| **课 时** | 8课时（360 min）。 | |
| **教学目标** | **知识技能目标：**  1．深入理解人工智能伦理的定义、内涵以及核心原则。  2．了解机器道德的基本概念、发展历程和理论框架。  **思政育人目标：**  让学生通过学习人工智能伦理，始终保持高度的伦理责任感和社会担当意识，关注技术对社会、环境、人类价值观的影响，积极倡导和践行伦理规范。 | |
| **教学重难点** | **教学重点：**人工智能信任危机的成因、表现及应对策略  **教学难点：**人工智能伦理规范 | |
| **教学方法** | 讲授法、问答法、讨论法 | |
| **教学用具** | 电脑、投影仪、多媒体课件、教材 | |
| **教学设计** | 第1节课：考勤（2min）--知识讲解（40min）--作业布置（3min）  第2节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min）  第3节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min）  第4节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min）  第5节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min）  第6节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min）  第7节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min）  第8节课：知识讲解（40min）--课堂小结（3min）--作业布置（2min） | |
| **教学过程** | **主要教学内容及步骤** | **设计意图** |
| **考勤**  **（2min）** | ■【教师】清点上课人数，记录好考勤  ■【学生】班干部报请假人员及原因 | 培养学生的组织纪律性,掌握学生的出勤情况 |
| **知识讲解**  **（40min）** | **【教师】**讲解遵纪守法—— 人工智能设计的伦理准则  **一、IEEE 发布《人工智能设计的伦理准则》白皮书**  美国时间 2017 年 12 月 12 日上午 9 时，电气电子工程师学会（IEEE）于全球发布了第 2 版的《人工智能设计的伦理准则》白皮书（“Ethically Aligned Design”v2），旨在指导我们认识这些技术可能造成的技术外的影响，确保人工智能设计能够符合人类的道德价值和伦理原则，并充分发挥系统益处。其所提供的洞察与建议，能为未来科技领域技术专家的工作提供重要参考，并推动符合伦理原则的国家与全球政策制定。其旨在指导人类伦理地设计、开发和应用 AI 技术，确保技术不侵犯人权，优先考虑人类福祉；同时确保设计者与操作者负责可问责，透明运行以降低滥用风险。  **二、《人工智能设计的伦理准则》的标准**  IEEE 所发布的《人工智能设计的伦理准则》主要包含以下十一个标准。  （1）解决系统设计中的伦理问题的建模过程。  （2）自主系统的透明性。  （3）数据隐私的处理。  （4）算法偏见的处理。  （5）儿童与学生数据治理标准。  （6）雇主数据治理标准。  （7）个人数据的 AI 代理标准。  （8）伦理驱动的机器人和自动化系统的本体标准。  （9）机器人智能与自主系统中伦理驱动的助推标准。  （10）自主和半自主系统的失效安全设计标准。  （11）合乎伦理的人工智能与自主系统的福祉度量标准。  **三、《人工智能设计的伦理准则》的意义**  IEEE《人工智能设计的伦理准则》明确要求，个人数据权利方面，人们有权决定个人数据的访问权限，可利用知情同意控制数据使用，并需机制帮助建立、维护独特身份和个人数据，同时明确融合或转售个人信息的后果。经济效应与福祉方面，IEEE 期望通过合理的通信网和互联网接入，使智能技术惠及各地人群，改变制度结构，促进其以人为本，并解决人道主义和发展问题，增加福祉。问责法律框架方面，由于智能系统与机器人技术融合，系统具有部分自主性及特定智力任务能力，甚至可能拥有人类外貌，因此需确保问责制，明确系统造成损害时的责任分配。  例如，智能与自主技术系统应适用相关的财产法，政府和行业利益相关者应该确定哪些决策和操作绝不能委托给这些系统，并制定规则和标准，以确保人类能够有效地控制这些决策，以及能够有效地为造成的损害分配法律责任。虽然自我完善的算法和数据分析可以使影响公民的决策自动化，但法律应该强制要求透明性、参与性和准确性，包括必须允许当事人、其律师和法院可以合理地获取政府和其他国家机关采用这些系统所产生和使用的所有数据和信息，如果可能，系统中嵌入的逻辑和规则必须对监管人员开放，并接受风险评估和严格测试，系统应当生成用于决策的事实和法律的审计数据，并服从第三方核查，同时公众有权了解是谁通过投资制定或支持关于这类系统的伦理决策。而有效的政策应当保护和促进安全、隐私、知识产权、人权和网络安全，以及公众对智能与自主技术系统对社会的潜在影响的认识。为确保政策最符合大众利益，这些政策应当支持、推广和实施国际公认的法律规范，同时提升劳动力在相关技术方面的专业知识，另外产生对研究和开发的引领作用，并制定规则，以确保公共安全和问责，最后教育公众知悉相关技术的社会影响。  IEEE 强调自主系统设计需注重伦理问题，特别是防伤害。要求武器系统具备人可控、可追溯、透明，操作人员需培训、可识别、可预测其行为。开发者需明后果、守伦理，慎处理有害开发。AGI 与超 AI 风险高，需防滥用、不良设计。  当系统接近并超过 AGI 时，无法预料的系统行为将变危险且难纠正，需谨慎并明确其运行机制。AGI 系统不一定与人类利益一致，能力增强时更需警惕。情感计算方面，智能技术需确保服务人类，避免人造物通过影响情感体验造成伤害，初步的人工情绪设计也需考虑其对决策者和公众理解的影响。混合现实技术普及后，可能改变身份和现实理解，感官增强设备的实时个性化能力可能引发个人权利和多重身份管理的伦理问题。 | **讲解遵纪守法—— 人工智能设计的伦理准则，让学生更加仔细的了解知识，从而激发学生的学习欲望。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了遵纪守法—— 人工智能设计的伦理准则，让学生了解IEEE 发布《人工智能设计的伦理准则》白皮书以及《人工智能设计的伦理准则》的标准。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  IEEE 发布《人工智能设计的伦理准则》白皮书的背景是什么？ | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解明辨是非—— 人工智能伦理问题  **一、算法歧视**  算法决策在很多时候其实就是一种预测，用过去的数据预测未来的趋势。算法模型和数据输入决定着预测的结果。因此，这两个要素也就成为算法歧视的主要来源。一方面，算法在本质上是“以数学方式或者计算机代码表达的意见”，包括其设计、目的、成功标准、数据使用等都是设计者和开发者的主观选择，设计者和开发者可能将自己的偏见嵌入算法系统；另一方面，数据的有效性、准确性也会影响整个算法决策和预测的准确性。例如，数据是社会现实的反映，训练数据本身可能是歧视性的，用这样的数据训练出来的人工智能系统自然也会带上歧视的影子。此外，算法歧视可能是具有自我学习和适应能力的算法在交互过程中习得的，人工智能系统在与现实世界交互过程中，可能没法区别什么是歧视，什么不是歧视。算法决策是在用过去预测未来，而过去的歧视可能会在算法中得到巩固并在未来得到加强，因为错误的输入产生错误的输出作为反馈，进一步加深了错误。最终，算法决策不仅仅会将过去的歧视做法代码化，而且会创造自己的现实，形成一个“自我实现的歧视性反馈循环”。归根到底，算法决策缺乏对未来的想象力，而人类社会的进步需要这样的想象力。  **二、隐私忧虑**  很多人工智能系统，包括深度学习，都是大数据学习，需要大量的数据来训练学习算法。数据已经成了人工智能时代的“新石油”。这也带来了新的隐私忧虑：一方面，如果在深度学习过程中使用大量的敏感数据，这些数据可能会在后续被披露出来，对个人的隐私产生影响；另一方面，考虑到各种服务之间大量交易数据，数据流动频繁，数据成为新的流通物，可能削弱个人对其个人数据的控制和管理。  **三、法律问题**  假设无人车伤了人，那么是开发者负责还是驾驶公司负责。关于人工智能带来的“伦理困境”，让很多专家感到纠结。假设出现如下场景：一辆载满乘客的无人驾驶汽车正在行驶，遇到一位孕妇横穿马路的突发状况，如果紧急刹车可能会翻车伤及乘客，但不紧急刹车可能会撞到孕妇。这种情况下，如果司机是人，那个瞬间完全取决于人的清醒判断，甚至是本能或直觉。无人车该如何选择呢？无人车依靠的是人工智能大脑，它目前不可能做超出人类算法中所设定范围的行为决策。将全国每年的交通事故数据“喂”给计算机，人工智能可以学习海量数据里隐含的各种行为模式。简单来说，就是无人车会从以往案例数据库中选取一个与当前情景较相似案例，然后根据所选取案例来实施本次决策。但遇到完全陌生的情景，第一个选择仍然是搜索，即在“大脑”中迅速搜索和当前场景相似度大于阈值的过往场景，形成与之对应的决断。如果计算机搜索出来的场景相似度小于阈值，即找不到相似场景，则算法会让无人车随机选择一种方式处理。  程序员可以通过代码来约定无人车如何做，但这种约定始终要遵循普遍的社会伦理。这个过程中，程序员要和伦理学家共同参与把关，程序员要将符合伦理的决策用代码的形式体现出来，输入系统中。以无人车为例，究竟由人工智能开发者负责，还是无人驾驶公司负责甚至任何的第三方负责，或者这几者在何种情形下各自如何分担责任，应当在相关人工智能的法律法规框架下通过制订商业合同进行约定。 | **通过教师讲解，熟悉明辨是非—— 人工智能伦理问题。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了明辨是非—— 人工智能伦理问题，让学生构建人工智能伦理框架，保障技术公正安全。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  探究抖音推荐系统的个性化与偏见。 | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解未雨绸缪—— 伦理风险评估（一）  **一、人工智能威胁用户隐私安全**  **（一）数据处理环节复杂，隐私安全风险高**  以数据为基础的人工智能产业链环环相扣、错综复杂，数据分析的采集终端、处理节点、存储介质与传输路径不确定，风险较高。鉴于人工智能产业链以数据为基础，环节众多且相互交织，基础训练数据作为核心生产要素贯穿其中。然而，数据分析的采集、处理、存储与传输过程充满不确定性，风险颇高，用户对这些环节缺乏必要的控制、监督与知情权。同时，行业内部尚未建立统一标准和行为规范，以明确相关企业在这些环节中的权责分配和行为合规性。因此，隐私数据的使用边界与安全保护实际上完全依赖于人脸识别服务提供商自身的安全技术能力、道德素养及行业自律水平。这种现状导致用户对自己的人脸隐私数据完全失控，进而可能使用户面临隐私透支的风险，并对现代人工智能技术产生抵触心理。  **（二）强人工智能产品可能诱导用户主动泄露个人 数据，挑战隐私边界**  随着人工智能技术的飞速发展，如何确保人工智能产品在规范框架下与人合理交互，不侵犯用户隐私并坚守职能界限，成为亟待解决的问题。除了底层算法模型训练数据和个性化服务采集的数据易遭遇泄露、篡改等风险，强人工智能产品可能突破设计伦理，主动或强迫用户提供非必要或高度敏感的隐私数据，尤其是具备自主意识与决策能力的教育机器人或同伴机器人。这些产品或在主观恶意或被操控下，引导、胁迫用户泄露敏感信息，用于非公开活动。因此，必须建立严格规范框架，确保人工智能产品与用户交互合理、安全，不越界侵犯隐私，这一命题值得我们高度重视。同时，还需加强监管和技术防护，确保人工智能产品的行为符合伦理和法律要求，保护用户权益不受侵害。  **（三）人工智能的滥用造成隐私透支，人格尊严更易被贬损**  以人工智能领域最常见的生物识别为例，面对处置用户生物识别信息命题，以苹果为代表的企业宣称自己仅将用户的指纹、人脸数据存储在终端设备本地，并采用物理加密的方式确保该数据的整个调用过程完全脱敏和本地化存储；但更多的厂商则是将人脸数据作为用户画像的一部分在线上随意流转传输，常见的网络传输与加密协议显然无法与人脸数据的安全级别与敏感程度相匹配，人脸识别应用的数据存储与传输流程均有可能被劫持，存在严重的安全风险；加之当前数据爬虫、网络入侵、数据泄露等已经成为互联网中的常态，收集人脸数据的供应商完全有可能主动或被迫在未经用户授权或超出用户协议许可的范围下，对用户面部数据进行采集、使用、流转等非法操作。  人脸特征信息一旦泄露，将带来严重的安全风险和隐私侵害。不法分子可能拦截或从被攻破的本地加密存储中复制这些信息，进而轻易攻破用户所使用的人脸识别服务，获取敏感权限如金融交易、人脸门禁、手机解锁及计算机敏感数据访问等。同时，这些人脸数据还可能被结合深度合成技术，伪造具有人格诋毁性质的多媒体内容，或捏造虚假音视频片段恶意侮辱、诽谤、贬损、丑化他人，对受害者造成极大的人格侮辱与内心创伤，并引发恶劣的社会影响，甚至被用于干预国家政治或执行军事行动。用户对人脸识别服务的依赖加深，使得隐私泄露事件的影响更加显著，而人脸识别的专属性与唯一性也会导致受害者难以挽回和消除由人脸数据泄露或非法篡改利用所造成的损失与不良影响。  此外，人脸作为人在社交活动中最直观的身份认证，难以通过有效手段施加保护，就可能导致用户在公共场合面临未经自己允许甚至不知情的情况下被非侵入性识别技术监控并抓取面部数据的风险，如 IBM 公司在被获取面部数据的自然人毫不知情的情况下从 Flickr 网站抓取近 100万张照片用于训练人脸识别算法，这无疑是对用户隐私的侵害。由此可见，人工智能的潜在隐私风险不仅会造成用户隐私权益遭受侵害，还有可能面临由隐私泄露导致的名誉权、肖像权、姓名权、信用权等权利遭受侵扰，其中的系统性风险不容忽视。  **二、人工智能侵犯用户生命财产等合法权益**  医疗机器人的广泛应用，虽带来了医疗技术的革新，但也伴随着诸多挑战与安全隐患。培育一名医生需十几年刻苦钻研与上千小时临床经验，知识架构与经验体系紧密相连，缺一不可。  而医疗人工智能产品的落地与商用，背后是千万量级数据、上百万行代码、几万份零件及众多供应商与生产线的复杂交织，行业背景、技术资源、产品判断标准等差异巨大。这不禁让患者对人工智能的安全性产生焦虑。如同人工手术需对主刀医师和医疗器械严格消毒，机器人手术则要求对机器整体无死角消毒。然而，鉴于机器的精密性与耐久性，我们难以确保对金属机器实施无死角消毒，也无法完全排除其他有害物质感染的风险，这为医疗机器人的应用蒙上了一层阴影。  即便从产业链角度来说万无一失，安全性得到保障，但鉴于人工智能厂商不会透露他们的人工智能系统如何工作，在任何一种情况下，当传统人工智能生成决策时，人类用户无法及时获悉该决策是如何生成的，尤其处于高危医疗环境，信任人工智能的决策而不了解人工智能给出建议可能存在的潜在风险根源，患者都会面临极大的风险。  手术机器人在执行高难度、高精确外科手术时，其操作过程中的任何微小差错都可能导致手术失败与患者死亡。然而，当这种差错发生时，往往难以界定责任归属：是现场操控医生的“直接人为错误”，还是机器人制造商、算法提供商的“间接人为错误”；是外部个人或组织的恶意物理攻击与远程干扰，还是电力、数据传输中断等意外事故；亦或是人工智能算法的黑箱特性或自主意识决策下的“机器错误”。这种责任界定的模糊性，使患者安全无法得到切实保障，也让人工智能在医疗（尤其是临床）领域的应用面临严峻的伦理挑战，步履维艰。  **三、人工智能动摇人公平参与社会活动，享有社会资源的权利**  人工智能产业链冗长复杂，涉及芯片、传感器、算法、终端、行业应用、解决方案、安全加密、网络传输等众多相关方，技术滥用与不正当竞争的风险难以根除。任何一方采取歧视性策略或不正当竞争举措，都将导致偏见与不平等传递至终端消费者，损害其公平权利。加之人工智能的硬件、算法、数据集等技术壁垒高耸，不公平现象一旦形成便难以打破。此外，机器学习的黑箱特性使得模型结果具有不可解释性，算法透明度备受争议。这不仅导致行为决策与训练数据之间的相关性难以精准评估，还可能被不法分子利用，进一步加剧阶层分化与种族特定群体歧视等问题。  例如，在医疗领域，尽管驱动 AI 赋能医疗产业（尤其是公共卫生体系）的初衷是解决医疗资源供求端矛盾，使现有医疗资源均衡化，国民健康管理结构化、常态化落实，但鉴于我国地区发展差异悬殊，如果现有医疗人工智能产业不够成熟，就激进推行医疗人工智能对现有公共卫生体系的渗透，AI 医学产品的消费零售价格仍居高不下，产品种类乏善可陈，产品性能与质量不尽如人意，偏远地区或中低收入群体消费者购买 AI 医疗设备的意愿就会大打折扣，不仅失去享有公共卫生资源的权利，更会因缺乏 AI 医疗相关配套设备，未形成数据闭环而在后续看病就医过程中遭遇更加负面的诊疗体验；或者为享受同等条件的医疗资源而付出更多的学习与资金成本，加重普通家庭的医疗负担。 | **通过教师讲解，掌握未雨绸缪—— 伦理风险评估（一）。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了未雨绸缪—— 伦理风险评估（一），让学生了解人工智能在收集、处理、存储用户数据过程中可能存在的隐私泄露风险。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  在人工智能日益渗透到人类生活各方面的同时，如何避免其过度干涉人类的自我意识与行为判断，保护人的独立自主性？ | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解未雨绸缪—— 伦理风险评估（二）  **四、人工智能干涉人类对自我意识与行为的判断决策，动摇人的独立自主性**  在特殊医疗场景下，人工智能的自主意识可能对人类个体的自我意识与行为准则产生威胁。例如，在健康管理领域，用于精神健康管理的聊天机器人，若判断失误传递错误信息，可能干扰交谈对象的正常自我意识，导致错误认知的形成。同样，在医疗保健领域，某些护理机器人可能通过过度保护的方式，如限制人身自由、减少积极活动，来“确保”监护对象的安全，这种做法实际上可能剥夺了监护对象的自主性与行为多样性。  **五、数字化劳动相关权利问题**  人工智能一方面对人类的主体地位造成影响，另一方面也在动摇人类数字化劳动相关权利。其主要体现在：对于劳动主体来说，人工智能打破了现有的劳动关系架构，威胁人类在生产活动中的主体地位；对于接受人工智能劳动生产服务的人来说，双方关系从某种程度上被异化。  **（一）劳动主体性地位问题**  劳动主体性地位问题逐渐深入生产、生活的人工智能，其优势令人类难以望其项背，人类在生产过程中的主体性地位受到挑战。  医生做出的每一个正确决策都基于扎实的专业知识和长期的经验积累，而人工智能强大的算力与信息储备能力，使其仅花费几秒钟即可完成一名主治医师需要花费十几年研习才能获得的全部医学知识和临床案例的学习。人工智能拥有比人类更高的精确度与工作效率，不会被情感与精力所限而拥有更小的出错概率，无论是对已有数据知识的回溯检索，还是对未来案例内容的学习更新，人工智能的速度都是人类无法企及的。尤其是针对流程烦琐或需要收集大量资料、分析大规模数据的工作，人工智能往往能够以最快的速度提供最优质的反馈意见和针对性的解决方案，这种高效、精准、便捷的模式一方面极大提升了人类的工作效率，在节省劳动成本的同时优化了工作成果，另一方面也威胁着人类对工作项目本身的决策与贡献权重，使人类职员遭受被取代的威胁和压力。除此之外，人工智能挑战人类在生产活动中的主体性地位，同时可能导致未来人类对人工智能技术的过分依赖而自我降低其对专业技能知识与工作经验的水准要求，甚至从潜意识层面开始动摇与质疑所学知识技能培养的存在合理性。  **（二）异化劳动者关系**  异化劳动者关系逐渐深入生产、生活的人工智能，导致人类在生产过程中的参与率下降，与人建立有效沟通与稳定关系的难度加大。  医疗人工智能的发展旨在通过模拟人类医师的智能行为，减少人为干预，提高诊断正确率和治愈率，同时减轻医生负担，提升医疗质量。随着医疗人工智能的深入应用，医护人员在医学检验等场景中的直接参与减少，患者与机器的接触增多。智能医学影像、自然语言处理等技术使医务工作者得以摆脱繁杂的基础信息收集工作，专注于疑难病灶的诊治。然而，医学诊疗不仅是科学与数据的碰撞，更是医患间情感的交流与支持。传统的直接沟通逐渐转变为借助人工智能设备的间接沟通，但医患间的有效沟通仍不可或缺。这种沟通不仅满足了患者抵御疾病未知焦虑的心理需求，也是医生给予患者心理支持的重要途径。有时，这种隐性精神需求与心理支持的达成，比药物或手术更能促进患者痊愈，体现了对医患双方的人格尊重与人文关怀。由此可见，一些特殊的职业其本质上体现的是人格尊重与人文关怀，以及绝大多数人依赖通过劳动行为建立对社会和人际关系的认知，人工智能的渗透可能从根本上改变这一现状，而其带来的影响仍旧值得深思。  **六、进一步加深数字鸿沟**  人工智能的应用与推广需要强有力的经济基础作为支撑，对于不发达地区和国家而言，智能技术所带来的对数字鸿沟的增大将严重影响其所能带来的福祉，同时，较差的信息基础设施也会使人工智能的发展捉襟见肘；对于发达地区和国家而言，大面积推广人工智能技术虽然具有足够的技术设施基础与经济基础，但城市中诸如老年人、残疾人等弱势群体由于其在教育、认知、生理等方面的特殊性，因此无法完全享受人工智能所带来的福利，这一点在城市数字化治理与基础设施建设中尤为凸显。  人脸特征信息一旦泄露，将带来严重的安全风险和隐私侵害。不法分子可能拦截或从被攻破的本地加密存储中复制这些信息，进而轻易攻破用户所使用的人脸识别服务，获取敏感权限如金融交易、人脸门禁、手机解锁及计算机敏感数据访问等。同时，这些人脸数据还可能被结合深度合成技术，伪造具有人格诋毁性质的多媒体内容，或捏造虚假音视频片段恶意侮辱、诽谤、贬损、丑化他人，对受害者造成极大的人格侮辱与内心创伤，并引发恶劣的社会影响，甚至被用于干预国家政治执行力军事行动。用户对人脸识别服务的依赖加深，使得隐私泄露事件的影响更加显著，而人脸识别的专属性与唯一性也会导致受害者难以挽回和消除由人脸数据泄露或非法篡改利用所造成的损失与不良影响。  技术所带来的数字鸿沟使他们在社会生活中边缘化，而数字鸿沟进一步导致权利分化；除此之外，随着技术的扩张，在技术风险不断膨胀的同时，技术福利却逐渐转变成基本规则；并非所有人都是技术的拥护者，每个人都有权力决定自己的人脸等隐私数据是否被用于其他用途，这就让那些希望降低技术风险或捍卫自身隐私数据而拒绝使用人脸识别的公民丧失了享有社会基本公共服务与基础设施的权利。 | **讲解未雨绸缪—— 伦理风险评估（二），让学生能够更好了解相关知识点。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了未雨绸缪—— 伦理风险评估（二），让学生了解人工智能干涉人类对自我意识与行为的判断决策，动摇人的独立自主性。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  简述异化劳动者关系。 | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解智防远虑—— 人工智能的信任危机与安全  **一、技术安全**  人工智能在带来科技感与便利性的同时，也伴随着可信问题。例如，特斯拉“自动驾驶”因无法识别道路清扫车而发生致死事故，聊天机器人经过训练后学会种族歧视，显然违背伦理道德和公平公正原则。达芬奇手术机器人因血液溅到摄像头而“失明”，感知系统失效。蓝牙音箱被监控导致失控，引发用户投诉。数据隐私问题也屡见不鲜，人脸识别公司数据库泄露等事件频发，数据被滥用迹象普遍，可能牵涉国家安全。这些安全隐患源于“内忧”如样本均衡、数据偏移、用户隐私，以及“外患”如对抗攻击、信息安全、场景受限。技术安全，包括数据安全、算法安全、网络安全等，均面临严峻挑战。  **（一）数据安全方面**  保护数据安全，是开展人工智能技术的基本必要条件。机器学习需要数量大、种类多、质量高的数据进行训练，从不同工程阶段来看涉及采集的海量原始数据、训练和测试数据集、应用系统现场采集数据等。数据使用存在大量安全隐患，包括采集安全隐患、使用安全隐患、共享安全隐患、传输安全隐患。  **（二）网络安全方面**  我国在互联网安全保障上经验丰富，但工业网络安全关注相对较少，尤其是与人工智能结合后的安全性。我国虽已积累大量互联网安全保障经验，并有专门队伍进行保障，然而对工业网络安全的重视程度仍显不足。以电网为例，美国国家电网曾遭受攻击，部分其他国家也出现过类似事件。随着人工智能与工业的深度融合，工业网络的安全性问题愈发凸显，亟需得到更多关注与重视。  **（三）算法安全方面**  算法的工作原理需要充分解析，如果算法正确，其结果便能够令人满意，而算法错误则会导致安全问题。在隐私安全方面，隐私安全的范围十分广泛，虽然从伦理道德的角度有诸多分析，但是并没有有效的办法解决隐私安全。算法模型是人工智能系统的核心，而算法模型中的安全隐患则可能给人工智能系统带来致命的安全后果。算法模型潜藏鲁棒性平衡、数据依赖等缺陷，如对抗样本攻击下的鲁棒性、数据集对模型准确性影响大、高可靠性挑战、数据质量问题、用户选择退出权难以保障。算法可能存在潜在偏见或歧视，即算法的潜在偏见或歧视导致结果偏差或处理不当“黑箱”特征，存在结果可解释性和透明性问题。深度学习在很多复杂任务上取得了前所未有的进展，但是深度学习系统通常拥有数以百万甚至十亿计的参数，开发人员难以用可解释的方式对一个复杂的神经网络进行标注，成为一个名副其实的“黑箱”。重要行业人工智能应用面临可解释性挑战。人工智能在金融、医疗、交通等攸关人身财产安全的重点行业领域应用时，人类对算法的安全感、信赖感、认同度可能取决于算法的透明性和可理解性。此外，除人工智能模型本身在技术上不透明性外，在数据、数据处理活动方面也可能存在不透明性。  **二、应用安全**  人工智能的应用安全主要涉及产品应用和行业应用两方面。产品应用，涵盖智能机器人、自动驾驶等按照人工智能技术对信息进行收集、存储、传输、交换、处理的硬件、软件、系统和服务。行业应用则是指人工智能产品和服务在智能制造、智慧医疗、智慧交通等领域的运用。这两者的安全隐患主要表现为由于人工智能应用是依托数据、算法模型、基础设施构建的，因此存在算法模型、数据安全与隐私保护、基础设施的安全风险。这些风险在人工智能应用中表现得更为突出，如攻击面更大、隐私保护风险更显著，特别是在自动驾驶、生物特征识别、智能音箱等领域。  **三、法律与伦理安全**  人工智能安全的第三大类是法律与伦理，涵盖法律法规、标准规范、社会伦理三个方面，目前在行业规范方面已经有深入研究，北京智源人工智能研究院联合北京大学、清华大学、中国科学院自动化研究所、中国科学院计算技术研究所等单位，共同发布《人工智能北京共识》，提出人工智能的研发、使用、治理，应该遵循有益于构建人类命运共同体和社会发展的十五条原则。人工智能的发展关乎全社会、全人类及环境的未来。准则对推进人工智能的研发、使用、治理和长远规划提出倡议。通过推动人工智能的健康发展，助力人类命运共同体的构建，实现对人类和自然有益的人工智能。  建立健全保障人工智能健康发展的法律法规、制度体系、伦理道德。人工智能的发展，意味着我们要面对一个更加丰富的世界，这个世界需要法律、规则和伦理的维护，确保人工智能安全、可靠、可控。例如，未来无人驾驶的普及需要道路交通安全法等法律做出相应调整；面临“救一个还是救五个”的伦理难题，智能指挥系统需要做出道德上的设定；用“大数据”收集、存储和分析个人用户信息，商业主体的行为需要被严格规制等。总而言之，在人工智能的研发和应用中，需要把握以人类价值观为导向的方法论，充分考虑人的良知和情感，避免出现安全失控、法律失准、伦理失常等问题。  在伦理道德层面，“智慧社会”强调“人”的主体地位，致力于构建新型的人机协同关系与伦理框架。随着“万物互联”时代的到来，万物似乎皆可通过连线、计算、识别码等技术紧密相连，主客体间的界限变得模糊。在这个人与物深度交融的世界里，丢失的物品能自主呼唤主人，窗户可感知风暴自动关闭，食物带有追踪识别码，冰箱能自动续订牛奶……面对这样的“人与物彼此嵌入”的新常态，我们更应坚守并提升“人”的主体地位，谨防“机器的生命化”与“生命的机器化”倾向，确保人工智能的发展始终遵循人类价值观，服务于人类福祉、社会进步与经济发展。  人工智能技术和系统应该是安全而稳健与可靠的，不应该轻易受到外部攻击的破坏。因此，需围绕上海建设具有全球影响力人工智能产业创新高地的总体目标，着力突破可信人工智能发展关键领域、关键环节“卡脖子”技术，在可靠性、可解释性、隐私性、公平性等方面探索通用人工智能理论、人工智能算法及系统可信性的验证技术，在芯片、软件、系统等方面推动自主创新，推动可信人工智能的产业发展，使智能科技更好地服务于产业发展与人民群众的幸福生活。  未来，应通过对人工智能相关法律、伦理和社会问题的深入探讨，为智能社会划出法律和伦理道德的边界，让人工智能服务人类社会。这也是世界范围内的一项共识，人工智能的发展应该以人类社会的稳定与福祉为前提。 | **通过教师讲解，熟悉智防远虑—— 人工智能的信任危机与安全。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了智防远虑—— 人工智能的信任危机与安全，让学生了解人工智能技术安全的基本原理和防护措施。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  在应用安全方面，特别是在医疗、交通等关键领域，如何确保人工智能系统的安全性和可靠性，防止因技术故障或误操作导致的严重后果？ | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解坚如磐石—— 可信的人工智能（一）  **一、G20 人工智能原则**  人工智能技术助力包容性经济增长，为社会与个体带来福祉。负责任地研发与应用人工智能，能有力推动联合国可持续发展目标的实现，成为驱动可持续、包容性社会发展的强劲动力，并有效降低对广泛社会价值构成的潜在威胁风险。  与此同时，人工智能与其他新兴技术一样，可能为社会带来挑战，如劳动力市场转型，隐私、安全、道德问题，制造新的数字鸿沟、人工智能能力建设需求。为了培养公众对人工智能技术的信任和信心，并充分发挥人工智能的潜力，G20 致力于以人为本的方法，该方法以经合组织关于人工智能的建议书《G20 人工智能原则》为指导。该原则还为决策者提供指导，以便最大限度地利用人工智能并分享其好处，尽量减少风险和问题，同时特别关注对发展中国家和弱势群体包容和国际合作。《G20 人工智能原则》包含两个部分，分别是“可信人工智能的负责任管理原则”和“实现可信人工智能的国家政策和国际合作的建议”。  **（一）可信人工智能的负责任管理原则**  1. 包容性增长、可持续发展及人类福祉  利益相关方应携手推进可信人工智能的负责任治理，共筑人类与地球的繁荣未来。应着力增强人类能力、激发创新活力、扶持弱势群体，致力于消除经济、社会、性别等一切形式的不平等，守护自然环境，推动包容性增长与可持续发展，全方位增进人类福祉。  2. 以人为本的价值观和公平  人工智能各参与方在整个人工智能系统生命周期中，应坚守法治、人权和民主价值的核心原则。这些原则涵盖自由与尊严的捍卫、自治与隐私的保护、数据安全的确保及不歧视原则的践行，同时秉持平等、多样性、公平与社会正义的理念，以及尊重国际公认的劳工权利。为实现这一目标，各参与方需立足当前技术水平与实际状况，采取有效机制与保障措施，确保人类自主决定能力的保留。  3. 透明度和可解释性  人工智能各参与方应对事关人工智能系统透明度的信息进行负责任的披露，信息应当有意义、切合实际并符合当前的技术水平。增进人们对人工智能系统的普遍认知；使利益相关方认识到他们在和人工智能系统进行交互，工作场所也不例外；让受系统影响的人理解人工智能系统产生的结果；对受到人工智能不利影响的那部分人，能够根据以简单易懂的信息呈现的（系统）因素，以及（系统）进行预测、建议或决策的逻辑基础，对系统的结果提出质疑。  4. 鲁棒性、信息安全性和物理安全性  人工智能系统需在其全生命周期内确保鲁棒性、信息安全与物理安全。无论是在正常使用还是可预见的其他条件下（如错误使用或不利环境），系统功能均需保持可控，并有效规避不合理的物理安全风险。为此，各参与方需确保系统的可追溯性，涵盖对数据集及决策流程的全面追踪，以便在技术和实际条件许可的范围内，对系统运行结果进行分析并回应质疑。同时，各参与方应依据自身角色、实际状况及能力，在系统的各个阶段持续采用风险管理方法，以全面应对隐私、信息安全、物理安全及偏见等各类风险挑战。  5. 问责制  人工智能各参与方应根据其角色、实际情况和当前技术水平，确保人工智能系统能正常运行，并尊重前述原则。  **（二）实现可信人工智能的国家政策和国际合作的建议**  1. 加强对人工智能研究和开发的投资  政府应当长期投资研究、开发及跨学科活动，同时鼓励民间投资；促进可信人工智能领域的创新，如集中力量解决技术问题及与人工智能相关的社会、法律、伦理及政策问题。政府还应当投资重要的开放数据集，同时鼓励民间投资，并尊重隐私、加强数据保护；建立一个有利于人工智能研究和开发的环境，防止非开放数据集存在的偏见因素；切实提高数据集的互操作性，促进使用共同标准。  2. 为人工智能搭建数字生态系统  政府应致力于构建可信人工智能的数字生态系统，涵盖数字技术、基础设施与知识共享机制等关键要素。为此，政府需积极推动数据信任机制的建立，确保数据在安全、平等、合法及伦理的框架下实现共享，为人工智能的健康发展提供坚实支撑。  3. 为实现人工智能赋能塑造政策环境  政府应当建立政策环境，促使可信人工智能尽快从研发阶段过渡到部署和运营阶段。为此，政府应建立试点，提供可控环境，对人工智能系统进行测试、升级。在适当情况下，政府应当审查和调整其现行的可信人工智能的创新及竞争政策规章框架和评估机制。  4. 针对劳动市场转型提升工人能力  政府需与利益相关方携手，共同应对社会就业转型的挑战。政府应致力于提升公众对人工智能系统的使用与交互能力，并提供全面的技能培训。同时，政府需强化社会沟通，确保工人在人工智能部署过程中能够平稳过渡。这包括为工人提供全职业生涯的技能培训，以及为被淘汰者提供保障性支持，助力他们重新融入新劳动市场。此外，政府应与各方共同倡导负责任的人工智能使用，确保工人的安全、职业质量得到提升，同时增强团队意识和生产效率，让所有人都能广泛、平等地享受人工智能带来的福祉。  5. 可信人工智能的国际合作  各国政府需与利益相关方紧密合作，共同推进负责任的可信人工智能部署。政府应在经济合作与发展组织（Organisation for Economic Co-operation and Development，OECD）框架或其他全球、地区性框架下，积极推动人工智能知识的共享，并致力于构建跨国、跨部门、开放多元主体的合作方案，以确保各国能够持续获得人工智能专业能力。同时，政府应倡导设立由多元主体参与、各方达成一致的全球技术标准，以支持互操作及可信人工智能发展。此外，政府还应鼓励建立评估体系，通过全面收集信息，来衡量各国在落实可信人工智能原则方面实际成效。  **二、欧盟人工智能原则**  人工智能作为当今时代最具变革性的力量，既为个人和社会带来了巨大利益，也伴随着一系列需妥善管理的风险。但总体而言，其收益远超风险。为确保人工智能的健康发展，我们必须坚守“最大化收益，最小化风险”的核心原则，并制订以人为中心的发展路径，明确人工智能的最终目标是服务人类福祉，而非单纯的技术进步。“可信人工智能”理念将引领我们前行，只有建立对技术的深厚信任，人类才能安心享受人工智能带来的全方位益处。  《欧洲人工智能战略》提出，人工智能应当坚持以人为本的价值观立场，人工智能本身并不是目的，只是服务人类的工具，其最终目的应当是增进人类福祉。据此，欧盟委员会认为，需要在现有监管制度的基础上制定伦理准则，供人工智能开发商、制造商、供应商和使用者在内部市场中予以采用，进而在所有成员国范围内建立通用的人工智能伦理标准。为此，欧盟委员会设立了人工智能高级专家组，并委托起草人工智能伦理准则。人工智能高级专家组于 2018 年 12 月发布了伦理准则初稿，随后与利益相关者、欧盟成员国代表进行会商。根据相关反馈意见，人工智能高级专家组于 2019 年 3 月向欧盟委员会提交了修订后的人工智能伦理准则。为了实现“可信人工智能”，本伦理准则确立了三项基本原则。①人工智能应当符合法律规定；②人工智能应当满足伦理原则；③人工智能应当具有可靠性。根据这三项基本原则和前述欧洲社会价值观，本伦理准则进一步提出“可信人工智能”应当满足的七项关键要求，具体包括：①人的自主和监督；②可靠性和安全性；③隐私和数据治理；④透明度；⑤多样性、非歧视性和公平性；⑥社会和环境福祉；⑦可追责性。虽然所有人工智能都应当满足这些关键要求，但是在实践中需要根据具体情况确定相应的实施标准。例如，用于阅读推荐的人工智能应用程序可能引发的风险肯定要比用于医疗诊断的人工智能应用程序引发的风险小得多，由此这两种人工智能应用程序对上述关键要求的实施标准也应当不同。  **（一）人的能动与监督**  人的能动性，作为以人为本价值观的核心，要求在人与人工智能共融的环境中，人的主体性必须得到保持。这意味着人工智能应成为增强人类自主性、保障基本权利的有力工具，助力个人基于自身目标做出更优选择，从而推动社会整体的繁荣与公平，而非削弱、限制或误导人的自主性。  为确保人工智能的健康发展，人的监督至关重要。这需要根据人工智能的特定应用领域，实施适度的人为控制措施，以维持人与人工智能之间的平衡。当人对人工智能的监督减少时，应相应增加对人工智能的广泛测试和严格管理，确保其行为始终符合人类社会的期望与规范。  人的监督可通过建立有效的治理机制来实现，如人机回环、人控回环等，这些机制能够确保人工智能在运行过程中始终受到人类的掌控和引导。同时，人工智能也需主动接受公共管理部门的监管，确保其行为不违背法律法规，不损害公共利益，从而保障人类社会的安全与稳定。  **（二）可靠性和安全性**  首先，“可信人工智能”要求其所用算法必须具有可靠性和安全性，完全能够应对和处理人工智能整个生命周期内其自身产生的各种错误结果。其次，人工智能具有可靠性和安全性，要求人工智能能够抵御来自外部的各种攻击和不当干扰，不仅能够抵御那些公开的网络攻击行为，也同样能够抵御那些试图操控数据或算法的隐蔽行为。最后，“可信人工智能”的决定必须是准确的，或者至少能够正确地反映其准确率，并且其结果应该是可重复的。  此外，人工智能应当确保每一步都具有可验证的安全性。这就要求，人工智能运行过程中发生的各种意外后果和错误都必须进行最小化处理，甚至在可能情况下进行可逆性处理。同时，针对人工智能运行过程中可能发生的各种潜在风险，应当事先予以充分披露，并建立相应的评估程序。  **（三）隐私与数据治理**  隐私和数据保护在人工智能整个生命周期的所有阶段都必须得到保障。根据人们行为信息的数字化记录，人工智能不仅可以推断出个人的偏好、年龄和性别，还可以推断出他们的性取向、宗教信仰或政治观点。为了使人们能够信任人工智能的数据处理，人工智能必须确保人们对自己的数据拥有完全的控制权，并且确保人们不会因为这些数据而受到伤害或歧视。  除了保护隐私和个人数据，高质量人工智能还必须满足一系列严格要求。首先，它应当采用高质量、无偏见的数据集，因为在数据收集过程中，人工智能可能会受到社会上偏见的影响，或纳入误差、错误，这些问题必须在训练前得到妥善解决，以确保人工智能的公正性和准确性。其次，人工智能必须保证数据的完整性，从计划、训练、测试到应用等每个步骤，所用方法和数据集都必须经过严格测试，并如实记录，以维护数据的可靠性和可追溯性。最后，高质量人工智能必须实施严格的数据访问管控，确保只有经过授权的人员和系统能够访问敏感数据，从而有效防止数据泄露和滥用，保障数据的安全性和隐私性。  **（四）透明度**  人工智能必须确保其行为具有可追溯性，如实记录并保存系统决策及其生成过程的所有细节，这包括数据收集的描述、数据标记的说明，以及所用算法的详细描述。同时，人工智能应提供其组织化决策过程的清晰阐述，系统设计选择的合理依据，以及相应的应用理由，以确保其数据和系统，乃至业务模型的全面透明度。此外，结合实际应用案例，人工智能的能力与局限性应以适当方式向各利益相关者进行充分披露。最后，人工智能必须具备可识别性，确保使用者能够明确知晓他们正在与哪个人工智能系统进行交互，并清楚该系统的责任主体是谁，从而增强人工智能使用的可信度和责任感。  **（五）多样性、非歧视性和公平性**  人工智能所用数据集（包括用于训练的数据集和用于实际应用的数据集）可能会受到无意识的偏见、不完整性和不良治理模型的影响。持续的社会偏见可能导致间接的或直接的歧视。当然，一些伤害也可能是人工智能开发商、制造商、供应商，甚至使用者故意利用偏见或者从事不公平竞争造成。诸如此类问题，在人工智能开发之初就应当予以解决。  为此，人工智能开发商在开发过程中应当建立多样化的设计团队，建立保障参与机制，这样有助于避免这些问题。同时，开发商不断咨询在人工智能生命周期内可能直接或间接受到系统影响的利益相关者，也有助于解决这些问题。另外，人工智能开发商应当全面考虑不同人群的能力和需求，确保人工智能具有易用性，并尽力确保残疾人也能够便利、平等地使用人工智能。  **（六）社会与环境福祉**  对于人工智能的影响，不仅要从个人的角度考虑，还应当从整个社会的角度考虑。人们应当认真考虑人工智能对社会公共治理活动的影响，特别是那些与民主决策过程有关的情形，包括意见形成、政治决策、选举活动等。此外，人工智能可以提高人们的社会技能，但同样也会导致人们的社会技能退化，因此人工智能对人们社会技能的影响也应当予以充分考虑。  为了确保人工智能具有可信任性，必须全面考量其对人类和其他生物的环境影响。理想状况下，人工智能的发展应确保所有人类，包括后代，都能享受到生物多样性和宜居的环境。这同样适用于当前全球关注的人工智能解决方案的可持续发展议题。因此，应当积极倡导人工智能在追求技术进步的同时，承担起可持续性和生态保护的责任，确保其发展成果惠及全人类及地球生态系统。  **（七）可追责性**  具有可追责性的人工智能能够有力提升人们对人工智能的信任。为此，应当建立责任机制，确保人们能够对人工智能及其结果进行追责。人工智能的内部审核人员和外部审核人员对人工智能开展的评估及相应的评估报告，是保障人们能够对人工智能进行追责的关键。因此，人工智能应当具有可审核性，并确保人们能够容易地获得相关评估报告。  人工智能的潜在负面影响需事先进行全面识别、科学评估、详细记录，并采取措施予以最小化。在面临需求间的权衡时，应以合理、公正的方式加以解决，并给出明确说明。对于那些影响人们基本权利，特别是涉及安全的重要人工智能应用程序，必须确保其具备外部可审计性，以保障其合规性和安全性。同时，针对人工智能应用过程中可能产生的不公平不利影响，应建立可预期、便捷且充分的补偿机制，确保受影响者能够得到及时有效的救济和补偿。  （1）鲁棒性，在人工智能的训练过程中部分场景可能没有出现过，因此对未知的情况应当具备一定的应对能力。  （2）自我反省，对自身性能或错误能够有所感知。  （3）自适应，当人类把一个通用系统应用到新的环境中时，便应当解决自适应的问题。如果没有自适应性，每次都需要做大量的适配工作，经济代价十分高昂。  （4）公平性，深度学习有时会带来决策上的不平衡，造成种族歧视，如机器被教会种族主义的例证。 | **通过教师讲解，认识坚如磐石—— 可信的人工智能（一）。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了坚如磐石—— 可信的人工智能（一），让学生了解并对比 G20 和欧盟提出的人工智能原则。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  面对日益复杂的网络环境，如何保障网络可信，防止人工智能系统受到网络攻击或数据泄露？ | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解坚如磐石—— 可信的人工智能（二）  **三、可信人工智能的关键问题**  人工智能算法的关键安全问题包括数据可信、算法可信、网络可信、应用系统可信。  **（一）数据可信**  当前，以深度学习为核心的人工智能技术与数据紧密相连，互为支撑。海量数据为人工智能的算力与算法带来了革命性的飞跃，推动了其快速发展。同时，人工智能技术也极大地提升了数据的管理、挖掘与分析能力，使得数据处理更为高效和精准。数据质量对于人工智能算法与模型的准确性具有决定性影响，直接关系到人工智能应用的安全性和可靠性。然而，人工智能算法强大的分析与决策能力若被恶意利用，特别是在数据采集与标注环节，将可能构成对人工智能系统的严重威胁，不仅会损害个人和企业的财产安全，更可能危及社会乃至国家的整体安全。因此，数据安全，包括数据治理、用户隐私保护、数据交易规范等方面，成为人工智能安全不可或缺的关键要素，需得到高度重视和有效保障。  欧盟 2018 年通过了《通用数据保护条例》，对个人数据与隐私数据的采集、管理与使用进行了规范与严格保护。为了进一步提升我国可信人工智能产业健康发展，急需在数据“采—存—用”过程中的采集存储、授权访问、共享发布、质量控制、隐私保护、标注安全、模型学习与智能决策等各个环节全面提升可信等级与开展技术突破。  **（二）算法可信**  算法是人工智能系统的大脑，定义了其智能行为的模式与效力。目前，以机器学习理论为基础的算法迅速发展，在图像、语音与文本等数据上检测、识别、预测等任务上取得了巨大成功，但由于目前其技术的局限性，对算法的分析仍停留在有限数据集上的准确率、召回率与计算效率等。因此，随着实际智能应用的推广，算法出现了许多性能外的可信问题。  人工智能算法的可信问题主要体现在内部与外部的安全隐患上。内部隐患源于算法本身，深度学习等典型算法在可靠性、公平性、鲁棒性、可解释性、透明性等方面存在理论与技术上的显著漏洞，这使得其决策结果具有不可预知性，且分析过程如同难以审计与溯源的黑盒，从而给  人工智能系统带来了安全隐患。此外，作为人工智能算法的载体，软件与传统可信软件理论之间存在理论与技术的鸿沟，诸如形式化验证等技术工具尚未被应用于人工智能算法与软件系统的可信验证中。外部隐患则主要来自新兴的恶意攻击手段，这些手段对人工智能系统构成了严峻的安全挑战。例如，对抗样本攻击能够通过精心设计的样本改变自动驾驶车辆或身份识别系统的行为；异常样本投毒则可能使智能系统产生错误的分类模式；而逆向攻击技术则可能通过大量的模型预测查询来窃取人工智能系统的模型。  **（三）网络可信**  网络作为信息化与智能化技术的支撑技术之一，其安全可信问题也是人工智能安全可信的关键要素。同时，人工智能技术也为网络安全注入了新的内涵。它涵盖智能安防、舆情监测、金融风控、网络防护等安全问题。在智能安防方面，国内摄像头的安全性令人担忧，较容易被攻破并进行数据造假，给用户带来利益和安全损失。在舆情监测方面，我国虽已积累很多经验，但是仍存在通过网络传播法律法规禁止的信息，组织非法串联、煽动集会游行或炒作敏感问题等非法事件，这将严重危害国家安全、社会稳定和公众利益。在网络防护方面，为了解决网络环境下面临的安全问题，保障网络信息安全，使用防火墙对网络通信进行筛选屏蔽，以防止未授权的访问进出计算机网络，按照一定的安全策略建立相应的安全辅助系统，采用人工智能技术对入侵检测、病毒攻击等进行行为分析与智能监测。  **（四）应用系统可信**  应用系统是人工智能应用在复杂动态“人—信息—物理”融合系统中的综合集成。因此，应用系统的可信问题表现在信息系统与物理世界的控制与交互、人在回路与人机交互及系统的整体性评估等方面。  （1）人工智能为智能制造、智慧交通、智慧电网、机器人等物理系统提供决策支持和控制模型，但以数据驱动的智能控制与生物控制在本质安全和理论基础上有较大差异。为了实现人工智能在物理系统上安全、可靠地运行，需要与人工智能可信决策相衔接的智能控制算法。实现人工智能系统与物理世界间控制与交互的可信是保证信息物理系统稳健的核心问题。物理系统中广泛存在各种强干扰，如风、光、电磁等，人工智能系统需要在干扰条件、不确定性物理环境下安全、稳定、可靠地运行；真实物理系统存在时延、饱和等强非线性约束，必须建立可满足强约束的人工智能稳健控制理论与方法。  （2）人工智能应用系统与人类活动空间发生深度耦合，如无人机、服务机器人等需要在开放空间与人类行为密切交互，此时一方面需要建立符合人机交互规律的人工智能算法与系统，另一方面，还需实现可保护人类安全的人工智能系统。同时，现有深度学习算法因可解释性差、鲁棒性弱等可信问题，在部分智能系统中无法提供本质安全的性能保证，故必须在“人—信息—物理”系统层面突破基于行为的人工智能可信理论与方法。  人工智能系统广泛涉及信息收集、交换存储、分析运算、芯片与硬件、软件与服务等多个环节，其在提供智能分析决策能力的同时，也面临着系统攻击面扩大、隐私保护风险加剧以及攻击效应长远化等问题。例如，在金融风控应用中，随着移动互联网与大数据收集分析技术的成熟，通过决策树、集成学习、置换分析等机器学习方法以及跨品类数据混搭运用、降维组合等手段，能够早期预判客户风险，为互联网金融借贷提供了有力支持。然而，与此同时，金融数据的泄露和客户信息的非法挖掘也带来了巨大风险，对移动互联网金融行业的健康发展构成了一定阻碍。 | **通过教师讲解，熟悉坚如磐石—— 可信的人工智能（二）。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了坚如磐石—— 可信的人工智能（二），让学生了解可信人工智能的关键问题。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  简述应用系统可信。 | 通过课后练习，使学生巩固所学新知识 |
| **知识讲解**  **（40min）** | **【教师】**讲解坚如磐石—— 可信的人工智能（三）  **四、可信人工智能的关键技术**  **（一）可信人工智能基础理论**  1. 人工智能可靠性  梳理“可靠人工智能”领域的科技发展方向（如人工智能系统高可靠形式化及测试验证、人工智能系统不确定性分析验证、安全攸关人工智能系统恶意攻击检测等），探索基础理论和关键技术，分别提出中长期及“十四五”的科技发展目标、思路和重点方向。积极推动人工智能在工控安全、电力系统、航空航天等国家支柱产业中系统可靠性设计、分析与验证等关键技术的突破，促进高端化、智能化产业升级。  2. 人工智能可解释性  梳理“人工智能可解释性”领域的科技发展方向（如黑盒人工智能系统可解释性，具备可解释性的人机智能模型框架设计和多学科（如心理学、认知科学、哲学等学科）驱动的可解释性等），挖掘关键技术、产业推动及社会民生需求。积极推动人工智能在医疗领域中影像分析、智能诊断与动态诊疗等基础理论与关键技术的突破，促进人工智能技术在医疗行业的落地。  **（二）可信人工智能共性技术**  1. 人工智能隐私性  梳理“隐私保护人工智能”领域的科技发展方向（如结合联邦学习、差分隐私约束下的统计学习、安全合规的建模技术、分布式贝叶斯学习、合成数据等技术的人工智能系统隐私加强技术、人工智能系统对个人数据的使用边界协商及建立等）。积极推动与人工智能相关的数据隐私、用户隐私、模型隐私等基础理论与关键技术的突破，促进人工智能在医疗、金融、征信与保险等高隐私性领域的技术革新与产业升级。  2. 人工智能公平性  梳理“人工智能公平性”领域的科技发展方向（如人工智能系统无偏见、群体或个体公平的算法策略、数据偏见的检测及度量、防止偏见增大的机器学习算法、多智能体强化学习中的公平策略、自然语言处理的公平性问题、非法偏见的人类准则引入等）。积极推动公平性在人工智能建模与评估等方面的基础理论与关键技术突破，促进人工智能在监控、风控、预测等重点领域关键环节的技术升级。  **（三）可信人工智能验证技术**  1. 人工智能系统可信度量  研究人工智能系统的可度量属性框架，以及相应的量化控制和度量评估方法；人工智能系统缺陷与可信性的内在联系、系统缺陷预测和缺陷分布规律；多维可信属性的多尺度量化指标系统、度量和评估机制及测评体系；可信属性间的交互关系及可能的涌现特征，包括多个属性或综合属性的相容与失配等；建立可信人工智能系统度量的技术标准或管理标准方案。  2. 人工智能系统可信预测  研究人工智能系统在环境和自身演化下的可信性质演化规律，以及系统在线演化的基础理论；基于系统行为的可信性增长和面向威胁的在线评估与预测理论；可信人工智能系统全生命周期的风险识别、评估、管理和控制模式及方法；人工智能系统中“环境—信息系统”的交互作用中的可信性质预测机制。  **五、伦理、法规与可信人工智能**  人工智能的创新和安全应是平衡发展，创新能够推动安全，没有创新，安全便没有意义。  在国际层面，已经有国家和组织提出可信人工智能的基本准则。欧盟提出的原则具体包括：福祉原则——向善；不作恶原则——无害，假设机器人与人类在同一工作岗位上，机器人出现失控并对人类造成伤害，责任主体该如何判定；自治原则——人类能动性，假设自主系统创造了独有的语言，人类无法读懂，安全性的保障至关重要；公正原则——公平，高新技术发展应当造福全人类，因此现阶段应不断完善通用人工智能的发展；可解释性原则——透明运行，人工智能的工作原理必须透明，这也是人类面临的困难和挑战。可信人工智能由我国率先提出，并在 G20 会议上获得了各国的广泛认可，成为当前人工智能发展的核心关注点。在人、机、物深度融合的当下社会系统状态中，如何对人、机、物间的物理运动及众多事件进行有效组合，并进行系统性的训练与测试，显得尤为关键。除此之外，推动人工智能的发展还需从系统规模、部件类型等多个维度入手，尽管这需要长时间的积累与投入，但技术的持续进步是不争的事实。在此过程中，应始终兼顾安全与可信，确保人工智能能够沿着健康的道路稳步前行，为人类社会带来更大的福祉。 | **通过教师的讲解，理解坚如磐石—— 可信的人工智能（三）。** |
| **课堂小结**  （3min） | 【**教师**】**回顾和总结本节课的知识点。**  **这节课我们一起学习了坚如磐石—— 可信的人工智能（三），让学生了解可信人工智能的关键技术以及伦理、法规与可信人工智能。** | 通过对所学知识的回顾，培养学生的归纳总结能力 |
| **作业布置**（2min） | **【教师】**布置课后作业  简述可信人工智能共性技术。 | 通过课后练习，使学生巩固所学新知识 |
| **教学反思** | 课堂营造更包容的课堂氛围，鼓励学生大胆表达，对积极参与的学生给予更多肯定。同时，建立多样化的反馈渠道。 | |