平 时 作 业 报 告

课程名称：计算机论题

学院：计算机与软件学院

专业：软件工程

指导教师：李俊杰

报告人：郑彦薇 班级：软件工程 01 班

时间：2022 年 4 月 20 日—2022 年 5 月 4 日

报告提交时间：2022 年 5 月 1 日

要求：

阅读材料《可解释 AI 发展报告 2022——打开算法黑箱的理念与实践》。请根据课本中第 1 章和第 2 章中有关伦理学和计算机伦理学的基本概念、研究方法、伦理学的分析方法等知识点以及课堂中所讨论内容，对人工智能算法公平性考量进行分析。具体要求如下：

（1） 什么是可解释 AI、并分析可解释 AI 与透明度、问责制之间的关系。（20 分）

（2） 结合第 1、2 章知识点，分析为什么要研究可解释 AI 机制？（20 分）

（3） 仔细阅读谷歌、IBM、微软等大厂的可解释 AI 机制，深入分析每个公司机制的使用场景。（20 分）

（4） 可解释 AI 机制融入于人工智能产品会对我们的生活带来哪些积极影响？请举例说明并详细叙述。（20 分）

（5） 报告写作：宋体、五号字体，不少于 1500 字；主要观点请用**粗体**标记；参考文献（如有）引用规范（20 分）。

说明：

（1）本次作业满分为 100 分，占总成绩的 10%。

（2）本次作业截至时间 2022 年 5 月 04 日（周三）23:59。

（3）报告正文：请在指定位置填写。

（4）个人信息：WORD 文件名中的“姓名”、“学号”，请改为你的姓名和学号；实验报告的首页，请准确填写“学院”、“专业”、“报告人”、“学号”、“班级”、“实验报告提交时间”等信息。

（5）提交方式：截至时间前，请在 Blackboard 系统中提交，延迟提交无效。

（6）发现抄袭（包括复制&粘贴整句话、整张图），抄袭者和被抄袭者的总成绩记零分。

（7）期末考试阶段补交无效。

（8）因版权原因，请勿在课堂以外传播本次作业所提供的阅读材料。

可解释 AI 相关问题探讨

（1）什么是可解释 AI、并分析可解释 AI 与透明度、问责制之间的关系。（20 分）

答：可解释 AI：可解释 AI 是为 AI 算法所作出的决策提供人类可读的以及可理解的解释；是确保人类可以轻松理解和信任人工智能代理做出的决策的一组方法，关注点在于对模型的理解，黑盒模型白盒化以及模型的可信任。

可解释 AI 与透明度的关系：在 AI 中，透明度是指对外提供 AI 算法的内部工作原理，包括 AI 系统如何开发、训练和部署，以及披露 AI 相关活动，以便人们可以进行审查和监督，保证信息的原始性和真实性。其目的是为相关对象提供适当的信息，以便他们理解和增进信任。透明度是可解释 AI 的主要目标，可解释性是实现透明度的一种方式。

可解释 AI 与问责制的关系：问责制是指让适当的组织或个人承担 AI 系统正常运作的责任的能力，问责制处于 AI 伦理层次结构的顶端，其他 AI 伦理目标都要为之作出贡献。确保能够对 AI 系统进行问责是可解释 AI 的核心目标之一。问责制是可解释 AI 的必要保障机制，可解释性主要是为了确保能够问责。

（2）结合第 1、2 章知识点，分析为什么要研究可解释 AI 机制？（20 分）

答：以深度学习为代表的新一代人工智能不断发展，而深度学习得到的模型是个黑盒，无法从模型的结构或权重中获取模型行为的任何信息，深度学习无可回避的难解释性和黑箱性促使人工智能的透明度与可解释性称为科技伦理的一个核心议题。从伦理学的角度出发，实现黑盒的可解释性，也符合了伦理抉择中的知情同意原则和社会公正原则。另外在材料中，也给出了可解释 AI 的意义--帮助用户增强对 AI 系统的信心和信任、防止偏见，促进算法公平、满足监督标准或政策要求、理解和验证 AI 系统的输出，推动系统设计的改进以及帮助评估风险、鲁棒性和脆弱性。

（3）仔细谷歌、IBM、微软等大厂的可解释 AI 机制，深入分析每个公司机制的使用场景。（20 分）

答：谷歌的 Model Cards 机制：模型卡片是谷歌公司推出的一项具有算法解释功能的技术，也是谷歌在可解释 AI 领域的最新实践成果之一。它是一种情景假设分析工具，能够为 AI 算法运作提供一份可视化的解释文档，能够为使用者阅读，使使用者更加充分了解算法模型的运作原理和性能局限，它在算法中扮演着与食物的营养成分表同样的角色。谷歌在其主页上提供了关于模型卡片应用的两个实例：人脸识别（面部检测）和对象检测。目前，谷歌的模型卡片的主要应用场景是谷歌云平台上的 Google Cloud Vision，这是谷歌推出的一款功能强大的图像识别工具，可以学习并识别图片上的内容，模型卡片为这种工具面部检测和对象检测功能提供了解释文档。

IBM 的 AI Fact Sheets 机制：IBM 研究院研发的 AI 事实清单是一项呈现算法模型重要特性的自动化文档功能，算法模型信息的标准化和公开化，有助于增强不同类型使用者对模型的理解和信任，并且能够以此避免因算法模型不透明而导致的系列问题。其功能是提供有关人工智能模型基本特征的重要信息，比如模型的目的、预期用途等等。以 AI 事实清单为代表的自动化文档是增强 AI 可解释性的重要方式，能够以一种清晰明了的方式，作为技术人员与使用者的沟通介质，从而避免许多情形下的道德和法律问题。AI 清单通过自定义的策略创建，自动、准确地从 AI 模型采集信息，并且自动生成报告文档，

有助于使用者为合规做好准备，降低管理风险。供应商自愿填充并发布事实清单，其中涵盖模型的关键信息，作为一种竞争优势，以保持自身在市场上的竞争力，也可以减少供应商和消费者之间的信息不对称。

微软的 datasheets for datasets 机制：数据集数据清单是主要聚焦于算法训练数据集的可解释工具。为每个数据集随附一个数据表，可以记录动机、组成、收集过程、推荐用途等；算法开发人员能够了解他们所使用的数据的优势和局限性，并且防范偏见和过度拟合等问题。数据集数据清单用于解决数据集创建者和数据集使用者这两个层面的关键需求。数据集创建者提高数据集的透明度，是使用者充分了解情况的必要条件；在充分了解的情况下，使用者能基于自身任务需求选择合适的数据集，并且避免无意间的滥用。数据集数据清单也可用于达到一些其他次要目的如：提高机器学习结果的可重复性，研究者和从业人员可以使用数据集中的信息重新调整或构造数据库而不用访问数据集等。

（4）可解释 AI 机制融入于人工智能产品会对我们的生活带来哪些积极影响？请举例说明并详细叙述。（20 分）

答：①在医疗影像上，可解释 AI 机制的融入提供了多维度临床辅助诊断参考信息，这样除了可以提升医生的诊疗效率，还能让医生在交互实践中逐渐增加对模型输出结果的信任度；同时提高医疗 AI 辅助决策模型透明度和增加模型输出参考信息多样化，能够更好的辅助医生通过眼底影像进行青光眼样眼底疾病临床诊断。

②在 CT 影像上，融入可解释 AI 机制的腾讯觅影软件提供了产品工作原理的详细描述，满足全局可解释性，帮助 AI 专业人员和产品用户（如医生）理解软件背后的模型特性，消除对因训练数据偏移而导致模型输出偏移的疑虑。

③在商业银行上，可解释 AI 的融入，增加了金融、理财等与用户财产安全相关的任务的透明度，增加用户对模型的信任度；另外，可解释 AI 的应用，也使得数据科学家和业务专家可以系统的监控和管理模型，从而不断优化业务成果，不断评估和改进模型的性能。

④随着越来越多的企业依赖人工智能提升自己的产品，不法分子也可以利用类似的能力实施大规模欺诈计划，可解释 AI 融入人工智能，增加它的透明度，可以使企业更加信赖新一代人工智能伙伴，也降低了不法分子在白盒进行违法操作的可能性。

写作部分：

对人工智能算法公平性考量的分析

在数字经济不断推进的大背景下，全球人工智能迅速发展，与多种应用场景深度融合。越来越多的人投入人工智能行业，它也逐渐成为推动经济创新发展的重要技术。深度学习算法的提出，使得人工智能技术应用取得突破性发展。2012 年以来，数据的爆发式增长为人工智能提供了充分的“养料”，深度学习算法在语音和视觉识别上实现突破。

人工智能技术的迅速发展，引发了一系列伦理问题。举两个例子说明：在教育应用中，人工智能的参与很大程度上影响了教师与学生的人文情感；人工智能应用企业和教育从业者一味地给学生提供更加智能的应用，学生在享受这些便利的同时，对智能应用的依赖也加深了。**人工智能教育应用还没有相应的技术应用规范和严格的法律法规制度来监管，其导致的伦理问题极大地阻碍了人工智能教育的进步。**在临床试验应用中，人工智能技术的发展，也渗透到了临床试验的各个领域，机器学习、深度学习等算法方法大幅提高了新药研发效率、临床试验的成功率。然而，由于临床数据收集方面缺乏监管框架，没有建立临床研究数据交换或访问网关等事实，把人工智能在临床医学上应用的

伦理问题暴露了出来。**与传统临床试验对比，受试者的数据安全和隐私保护是一个新的挑战，信息技术的代替，也使得越来越多的电子知情同意代替传统纸质知情同意，但电子知情同意使用过程中，对受试者信息的采集、电子签名等的合法性仍存质疑，知情同意的合法的道德仪式可能被削弱而损害受试者权益。**

人工智能的关键所在是算法，**在技术语境下，人工智能算法可以被称为一种通用算法。这意味着，与传统计算机算法相比，人工智能算法更多地呈现了一种总分关系。**对于**人开发设计的算法，很难保障其准确性和无歧视性，人工智能算法的输出具有不确定性，长时间应用容易形成一定的歧视和偏见**。我们是否可以根据算法分析数据的结果给出黑人犯罪率高于白人的定论？当这样的分析结果作为人工智能算法的一个收获公布出来时，难免成为一次赤裸裸的、公然的人种歧视，甚至爆发一次激烈的争论、引发不必要混乱。算法结果必然伴随着人种差异、性别差异、年龄差异、地区差异等等问题，这无疑在一定程度上会被视为一种使用科技成果、利用数据分析而进行的一次人种、性别、地区歧视。**让人工智能算法的开发设计“暴露”于大众眼前，对于人工智能为大众所接受、所信赖，是至关重要的。**

可解释 AI 的提出，使得**人工智能算法更加透明化，然而在不足的监管力度面前，也容易出现一些差别化对待的现象。**例如企业为实现利益最大化，可能滥用算法技术，出现我们常说的“杀熟”现象：企业利用大数据计算顾客的行为偏好，针对同一产品给出不同的价格。

一味的“公平”不可取。例如在审批贷款时，贷款人需要根据借款人的基本情况、信用等级评估等具体信息签订不同的贷款合同；又如利用人工智能算法辅助刑罚时，需要根据案件性质、被告人身份如未成年、孕妇等给出不同的判决。这些领域的特殊要求，意味着**对于人工智能这样的通用算法，在解决其结果的不确定性、歧视或偏见的问题时，不能采取一刀切的方式。**

人工智能算法的公平性不仅仅在于“无差别”提供商品价格、“无差别”提供医疗条件、“无差别”对待地区文化差异等等，还在于对于不同身体状况的人提供不同的医疗、针对不同的要求和信用提供不同的贷款服务、对不同的刑事案件给出不同的刑罚等等。另外，**在对人工智能法律责任制度安排时，人工智能是否应拥有法律地位？若赋予人工智能主体资格，又如何对其法律责任进行规定和解释？**我想这也是在考量人工智能算法公平性时应考虑的问题。

对人工智能算法公平性的考量，实质上是对于人工智能算法如何进一步规制、如何针对事实制定人工智能算法问责制度等关于人工智能发展的提问。可解释人工智能对于人工智能发展的益处，使得可解释 AI 的发展以及其融入生活的各个领域成为必然，是科学技术发展的必然结果。人工智能的快速发展，使得相关文件与政策的出台刻不容缓。

参考文献：

[1] 刘星、卢晓然、吴影等. 人工智能应用于临床试验的伦理问题分析及对策. http://www.cjcpt.com. 2022 Mar;27(3):322-327.

[2] 刘艳红. 人工智能的可解释性与 AI 的法律责任问题研究. 法制与社会发展（双月刊），2022 年第 1 期.

其他（例如感想、建议等等）。

为了完成本次大作业，搜索了很多关于人工智能、可解释人工智能的研究。对人工智能可能引发的伦理问题，可能造成的不合法行为有了更全面的认识。阅读完相关文献，

也更加认为任何一项科技的发展过程中一定伴随着各种各样的问题，如何使这把双刃剑更好的应用于人类生活，是科技发展的共同目的。

指导教师批阅意见：   成绩评定：   指导教师签字： 年　月　日
备注：