**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： 董延杰**

**报告人： 谢弘烨 学号： 2020161036 班级： 软工02**

**时间： 2022年3月23日—2022年4月7日**

**报告提交时间： 2022年4月6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4月07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）**  可解释AI（Explainable Artificial Intelligence，即XAI）是指系统自身各个算法模块的输入、输出和性能如何促成系统结果以及这四者各自均可以理解，并提供阐释说明的人工智能系统。  对于可解释AI的讨论离不开透明度和问责制。其中，透明度的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；问责制是指让适当的组织和个人承担AI系统正常运作的责任的能力。简而言之，透明度是可解释AI的主要目标，而可解释性是实现透明度的一种方式；问责制是可解释AI的必要保护机制，而可解释性主要是为了确保能够问责。同时，AI的透明度、可解释性本身不是最终目的，而是实现其他目的（如问责）的手段和前提条件。  **（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）**   1. 传统AI所具有的“黑盒”特性，已经**涉及到伦理学研究的基本问题，即道德与利益两者的关系问题。**部分个人与团体可能以AI的“黑盒”性为借口，做出违背道德的行为以牟利，可解释AI可以有效避免这样的情况发生。 2. 不可解释的传统AI已经违反**网络伦理中的知情同意原则**，如用户无法得知AI系统的某项输出结果的具体计算过程；影响到**社会公正原则的具体践行，**如在厂商侵犯用户权益时，AI的运行过程无法公示以作为司法证据。   **（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）**   * 谷歌：模型卡片（Model Cards），以通俗易懂的方式展示了现实算法的基本性能机制与显示算法的关键限制要素，进而辅助用户更好地审视和理解算法的运行路径，确保用户对算法决策的知情和理解。**使用场景**：尚未得到大规模应用，谷歌仅提供了两个应用实例“人脸识别（面部检测）”和“对象检测”。在人脸识别实例中，模型卡片首先提供包括输入（照片和视频）、输出（检测到的每个面部机器相关信息）在内的“模型描述”，即算法基本功能。除此之外，模型卡片还通过提供“算法成分表”的方式，向用户展示算法的运行原理，对于不同数据的性能和局限所在。其目的是帮助开发人员就模型的选择及部署做出更明智的决定。模型卡片的应用场景远不止此，**对于专业人员，模型卡片可以用于更好的了解算法的性能和局限并以此改进优化算法；对于行业分析师与媒体，模型卡片可以用以向普通大众更好地解释算法的复杂原理，进而使普通人能更好地从算法的透明性中获利。**总之，它不光能促进技术人员之间的对话，还能促进专业人士与普通人的对话。 * IBM：事实清单（AI Fact Sheet），一项标准化、公开化呈现算法模型重要特性的自动化文档功能，包含了策略创建（允许用户自定义所要展示的的AI信息已经所要记录信息的相关环节）、自动数据采集（在AI系统整个生命周期中连续自动地采集数据）和自动报告三大主要功能，有助于增强用户对算法的信任，避免因其不透明性产生的系列问题。**使用场景：**极高的个性化程度是AI事实清单的最大特性之一，其收集到的信息可由行业、领域、数据、用例、用户、级别不同而各不相同，同时又能满足不同的透明度与合规要求。不仅如此，AI事实清单所展示出的信息类别多样且详细。因此，它**不仅可以加快开发人员对AI伦理的认可与接纳，还能够提高AI系统的透明度，增强用户对人工智能的信任。**除此之外，由于在最初的设计中，AI事实清单还以“供应商合格声明”（SDoC，Supplier’s Declaration of Conformity）为基础原则，**研究者设想其可由AI服务供应商自愿提供以保持自身在市场上的竞争力，进而促进整个AI服务行业良性发展，减少供应商与消费者之间的信息不对称以赢得消费者的信任。**另外，AI事实清单创建过程中无法删除已添加的信息，使得在对算法优化过程中的任何改动都将创建新版本的事实清单，**进而便于开发者与旧版本算法进行对比，以更准确地监视已部署的服务。**在AI事实清单之后，IBM还推出了包括AI Fairness 360，AI Explainability 360和Adversarial Robustness Toolbox等一系列可解释AI的开源工具包，进一步推动可解释AI的行业发展。 * 微软：数据集数据清单（Datasheets for Datasets）是主要聚焦于算法训练数据集的可解释工具。它为每一个数据集建立了一个对应的“数据表”，其中包含了数据集的动机、组成、收集过程、推荐用途等详细数据，便于开发人员了解所用数据集的优势与局限性，加强数据集创建者与使用者之间的沟通与交流。**使用场景：对于数据集创建者，数据表可用于鼓励对数据集的创建、分发和维护过程进行思考与检索，其中包括可能的风险与危害，并着力提高数据集的透明度；对于数据集使用者，数据表可确保其获得所需的信息，并以此指定有关数据集使用的恰当决策。**此外，数据表还可用于提高机器学习结果的可重复性，便于研究者与从业人员调整、改变数据集。   **（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）**  对于带有辅助驾驶功能的车辆，如果出现交通事故，可解释AI可以提供车机在事故发生时刻所做出的决策，从而能够为执法部门提供责任划分的重要证据。  手机、平板、电脑等个人电子产品中所带有的人工智能系统在可解释AI机制的融入下能够更好地被用户所使用。  执法部门在融入可解释AI机制的人工智能辅助工具的帮助下可以将原来模糊无法作为证据的视频、音频修复来作为指控犯罪嫌疑人的作证。  **其他（例如感想、建议等等）。** |

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |