**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： 计算机科学与技术**

**指导教师： 杨艳丽**

**报告人： 刘俊楠 学号：2017303010 班级： 01**

**时间： 2021年3月23日—2021年4月9日**

**报告提交： 2021年4月7日**

|  |
| --- |
| **要求：**  阅读材料《人工智能算法歧视及其治理》和《一个都不能用？62个AI算法被指存在重大问题，剑桥团队：都不具有新冠临床诊断价值》。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. 什么是人工智能算法？人工智能算法公平性问题是如何产生的？请根据《人工智能算法歧视及其治理》和《一个都不能用？62个AI算法被指存在重大问题，剑桥团队：都不具有新冠临床诊断价值》展开叙述人工智能算法公平性问题产生的原因。（20分） 2. 人工智能算法公平性在日常生活中的存在形式有哪些，这些存在形式有什么样的特点？（20分） 3. 如何应对“不公平”的人工智能算法，请列举至少两个应对方法并详细叙述。（20分） 4. 公平性融入于人工智能算法设计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。（10分） 5. 你认为人工智能算法的公平性是利大于弊还是弊大于利，请选择一个持方进行详细论述（论点和论据可不局限于阅读材料）。（10分） 6. 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2021年4月7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。 |

|  |  |
| --- | --- |
| **人工智能算法公平性问题的成因及应对策略**  **（1）什么是人工智能算法？人工智能算法公平性问题是如何产生的？请根据《人工智能算法歧视及其治理》和《一个都不能用？62个AI算法被指存在重大问题，剑桥团队：都不具有新冠临床诊断价值》展开叙述人工智能算法公平性问题产生的原因。（20分）**  答：  人工智能算法：所谓人工智能算法（AI Algo-rithm）指的是：**“在计算机科学中用于描述一种有限、确定性和有效的问题解决方法，适合作为计算机程序来实现。”**  人工智能算法公平性问题指的是：**算法在数据集构建、目标制定与特征选取、数据标注等环节中产生的信息偏差，导致算法失去公平和准确性，** 主要表现为年龄歧视、性别歧视、消费歧视、就业歧视、种族歧视、弱势群体歧视等现象。[1]ＡＩ算法应用 正变得越来越普遍，越来越多的利益分配和大数据 直接相关，尤其是在算法决策应用日益广泛的教育、 就业、福利补贴发放、刑事司法、公共安全等重要与高价值领域，算法歧视可能会导致严重的政治与道德风险。  算法公平性问题产生原因：  **第一，算法研发者的偏见。**算法及其决策程序终究是由人设计的人造物，研发者的利益与价值取向难免会嵌入其中。由于算法黑箱的“阻隔”，普通人看到的只是结果而非决策过程，很多人在毫不知情的情况下承受着种 种隐形的歧视与精准的靶向不公正。  **第二，样本与训练数据的偏见。**样本与训练数 据通常被喻为算法的“教科书”，机器学习的效果与样本平衡与否、训练数据多少密切相关。训练样本不能太少，各个类别的样本数量差别不能太大，太少、差异大的数据无法有效代表数据的整体分布情况，容易造成过拟合。  **第三，算法研发公司以及购买企业的利益诉求。**算法研发的最终目的还是为了应用牟利，因此，研发公司与购买公司的意图无疑会深刻地反映在算法设计中。如果某些企业把利润追求凌驾于企业社会责任与社会公德之上，算法歧视也就难以避免。  **第四，算法自身原因造成的歧视。**主要表现为四方面：一是，算法黑箱，自我解释性差。二是，算法极其复杂，涉及大量专业知识，难以理解，即使专业人士很多时候都未必能在短时间 内了然其设计结构，对于普通人来说，若想洞悉某一种算法的奥秘，难度可想而知。三是，目前，人工智能领域陷入了概率关联的困境：不问因果只关问相关性，只做归纳不做演绎。四是，算法已经自我学会歧视，也容易被“教坏”。  **第五，西方中心主义意识形态也是算法歧视普 遍存在的重要原因之一。**世界上最强大的互联网公司以及最先进的算法技术大多集中于西方国家，因此，西方国家根深蒂固的西方中心主义意识形态难免以“内隐性社会认知”的方式嵌入算法之中。尽管西方国家高举种族平等大旗，但实际上，基于西方中心主义的无形的偏见与歧视却根深蒂固。[2]  **（2）人工智能算法公平性在日常生活中的存在形式有哪些，这些存在形式有什么样的特点？（20分）**  人工智能算法公平性问题，在日常生活中表现的形式有:  1. skewed 样本。**如果初始使用的是有偏的样本，这些偏见可能会随着时间一直存在。**未来的观测会导致更少的机会来纠正预测。例如警察发现某个地区犯罪率高，会更加关注这些地方，更容易记录这些地方的犯罪，后面即使其他地区犯罪率也增高，但是可能由于警察的关注比较少，所以记录的犯罪率比较低。  2. **大数据会记录你的浏览习惯，购买习惯，常用淘宝支付宝这些软件的人，你的消费能力、购物习惯、活动产所、收入情况、生活质量、年龄、身高、体重、鞋码、三围、口味等**，都是可以分析出来的 这些基本囊括了我们的生活。个人数据安全就成了一个大问题，一旦数据泄露（或被买卖），可能会对用户人身财产、国家和公司的安全造成威胁。各类软件，在未经用户许可之下采集个人信息. 据央视报道称，苹果的定位功能不仅可以知道用户的家庭住址、工作单位，甚至每天什么时候去哪儿，待了多长时间，都会被精准地记录下来。还有用户在手机上使用软件所在的位置，也都会被记录下.即使关了这个功能，后台依然会把位置信息记录下来。[3]  3. 污染样本。**很多机器学习系统中存在一些认为造成的偏差。如果一个简历筛选系统使用管理员的label来训练模型而不是根据申请者的能力。**那么这个系统训练得到的模型会更倾向于选择管理者的决定。  4. **杀熟，即同样的商品或服务，老客户看到的价格反而比新客户要贵出许多。**近两年来包括滴滴出行、携程、飞猪、京东、美团、淘票票等多家互联网平台均被曝疑似存在“杀熟”情况，涵盖在线差旅、在线票务、网络购物、交通出行等多个领域，特别是OTA（Online Travel Agent）在线差旅平台较为突出。[4]  第一种形式容易造成地域，教育，以及人种差异带来的歧视，这种形式实际上是**现实世界存在的歧视的延伸**。  第二种形式通过对用户信息的精准采集，了解用户的消费习惯、物质需求，企业可轻易**实现对不同消费群体歧视性提价**。**这种形式防不甚防**. 大家在享受生活便利的同时，却开始担心个人隐私安全。而一些信息泄露引发的骚扰诈骗的后果，也让用户人人自危，害怕自己一个不经意的举动就会让信息暴露在空气中，成为不法分子行骗的“诱饵”。  第三种形式会造成部分群体**难以就业**，并且这种形式难以证明，应聘者难以找到相关证据来维护自身的权益。  第四种形式对于消费者个人来说，一次交易中或许仅仅是价格的小数额差异，但当消费者对平台产生依赖，长期而言交易成本不降反升，侵害了消费者的公平交易权；对于全社会而言，**杀熟行为破坏了诚实信用的基本原则，降低消费者的消费欲望，导致全社会交易成本的上升。**  **（3）如何应对“不公平”的人工智能算法，请列举至少两个应对方法并详细叙述。（20分）**  首先，政府监管方与行业专家，根据行业需求制定合理的公平性准则，并制定算法歧视的问责法律。  1、可以**通过立法，制定相关的规范来约束**,确立适合国情的“综合保护”立法模式.  欧盟于2012年提出了《欧洲议会和理事会关于个人数据处理中个人权利保护及促进个人数据自由流通条例草案》，欧盟统一立法模式可为我国提供有益借鉴，应鼓励和引导行业自律。[5]  2、**加大行政处罚的力度是当务之急。**  权利保障与救济是个人数据保护立法的核心。在云计算模式下，侵权主体具有多样性并很难确定，数据主体难以有效维权。我国应逐渐健全个人数据保护的民事诉讼机制，明确建立由被告证明其不应承担责任的举证规则。当数据主体不能确定具体的侵权主体时，由数据控制者与处理者承担连带侵权责任。此外，在大数据时代，个人数据是价值巨大的“金矿”，很多企业会对存储于云端的数据进行商业化利用，数据泄露、非法交易也将成为常态，可考虑**设立违法行为处罚“公示”制度，将违法企业计入“违法行为黑名单”等方式**，来有效遏制侵犯个人信息的违法行为，以及算法实现的功能，由公众来判断是否合乎社会公平的需求。**通过群众的判断**，算法研究者更容易看到设计的算法机制中存在的问题。  3、**与欧盟或相关国家展开双边磋商，建立一个类似于欧美的“安全港协议”。**  跨境的数据流动将成为常态。中国仍被欧盟认定为是不能对个人数据提供充分性保护的国家，这将使具有财产价值的个人数据外流易，而流入难，导致企业在国际贸易中始终处于竞争劣势，甚至我国大量公民个人数据被变相获取。应对这一问题的最主要方法是**与欧盟或相关国家展开双边磋商，建立一个类似于欧美的“安全港协议”。同时，积极参与到个人数据保护的国家合作中，争取规则拟定中的话语权。**在国际立法中，应强调将个人数据流动作为经济问题的解决方案，而对个人数据保护在人格权方面的问题适度搁置，以利于纠纷的解决。  4、**完善“通知—同意”法律规则**  **大数据背景下的“通知—同意”法律规则难以发挥应有的功效。为应对此难题，欧盟的做法是明确数据主体的“同意”应为“积极同意”，**这有利于保护个人数据权利，但也因增加商业成本而饱受诟病。建议我国在数据保护立法中采用“积极同意”与“消极同意”相结合模式。对于统一的数据保护立法应采纳“消极同意”模式。考虑到这种模式所降低的商业成本会变相转移到数据主体身上，故我国在引入消极“通知—同意”规则时应注意两个因素：一是“通知”应是明确、充分、具体的。数据控制者应告知数据处理的机构、目的、种类等核心要素。二是赋予数据主体便捷、经济的反对途径。应规定数据处理者要求数据主体反对的方式必须与数据处理者通知的方式同等便利，不给数据主体增加额外的负担。对于犯罪侦查、医疗、科研等特殊领域或涉及到个人敏感数据的特殊保护规则应采用“积极同意”模式。  5、**科研开发者在此基础上通过设计实现算法公平。**在算法设计之初，就将算法公平性准则、算法可解释性、算法问责等价值需求囊括在算法设计之中，这也督促设计者在设计阶段严格遵守公平性的伦理和法律规则。  6、在算法的应用阶段，**政府监管方与用户，共同监督算法公平性实施。**对严重的算法歧视行为，由政府监管方进行问责。通过各方联合起来，解决算法的不公平问题，人工智能才能够被放心地应用于民生中的各个领域，并真正地造福全社会。  **（4）公平性融入于人工智能算法设计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。（10分）**  1.消除因性别，年龄，人种上的歧视带来的网上求职机遇不同。  融入公平性，可以让不同年龄，人种，性别的应聘者在同一个平台上公平竞争，一方面对于应聘者，不会因为一些外来差异而失去工作机会；另一方面对于公司，不会导致人才流失，公平性的融入可以让公司有机会引入更多的人才。  2.融入公平性，减少不同人种犯罪率的巨大差值，维护社会稳定  通过融入公平性，降低对高犯罪率人种的再次犯罪预估，给予更多机会以参加学习教育，一方面不会导致不同人种巨大的犯罪率预估值差距带来社会矛盾，一方面高犯罪率人群也有机会接受教育，从源头改变社会现状。[2]  3.保障消费者的权益  通过公平、一致、透明的机制，可以保障消费者享受同等的服务。避免各种软件通过对消费者的信息收集对不同消费群体差别定价的操作。    **（5）你认为人工智能算法的公平性是利大于弊还是弊大于利，请选择一个持方进行详细论述（论点和论据可不局限于阅读材料）。（10分）**  **我认为人工智能算法的公平性是利大于弊。**  1.**人工智能算法中隐藏的歧视比现实社会中的更可怕。**  **算法中隐藏的歧视不容易甚至不可能被发现**，但它又时时刻刻作用于我们的生活，这正是其可怕的地方，就像《人工智能算法歧视及其治理》中所论述的“被嵌入种族歧视代码的算法中隐藏的‘歧视特洛伊木马’，在人工智能‘客观、公正、科学’的高科技包装下更容易**大行其道**，在算法黑箱的遮掩下更隐蔽。”在算法中嵌入歧视，更容易潜移默化的影响更多人，**歧视问题将成为更深入，更顽固的社会问题**，滋生更多的伦理道德冲突。  2.**在算法中融入公平性，能让企业和员工最大利益化。**  通过消除公司对工人简历而采取的在性别上，种族上，年龄上以及身体上有差异的筛选的方法，能够让每个员工更多地凭借个人工作的经验与能力通过公司的面试。这些对于员工，不至于因为莫名其妙的原因失去理想的工作机会，对于公司来说，不至于自己想要的人才白白流失。**若采取一定的门槛限制，一方面会引起民众的不满，另一方面公司会有很大程度的损伤**，例如阅读材料中雷诺烟草公司根据年龄筛选在线申请人、但未向任何被拒绝的候选人透露此事的丑闻被举报的事件，既影响了申请人的未来，也对公司造成了不好的影响。[2]  3.**算法公平性的引入，有利于保护弱势群体。**  在歧视性的算法机制下，根本无法保障弱势群体享有同等的车贷，房贷，保险，教育，医疗等应有的福利待遇。对于弱势群体的数据采集分析，也会让一些黑心企业有机可乘；**对于缺乏教育者，生活困难者等更容易被差别对待或遭遇网络欺诈**；导致诈骗多样化。例如在材料中，美国联邦贸易委员会调查发现：广告商倾向于针对生活在低收入社区的群体推送高息贷款信息，这将导致更加严重的社会问题。因此算法公平性是利大于弊的。  **其他（例如感想、建议等等）。**  每种定义算法公平性的方法都有其优缺点。**我认为最令人不安的不是每种方法所面临的弱点，而是这些方法从根本上说是互不兼容的。**当使用受保护类别作为检测算法公平性的基线时，我们不能忽略受保护类别。我们不能要求相似的算法错误率，同时要求相似的风险评分在群组间必然会产生相似的结果。定义算法公平性的竞赛还在进行中。民主党人、共和党人和自由主义者对什么是算法公平性无法达成一致，而我也认为把算法公平性当作一个数学和计算机科学问题来对待有点过于乐观了。问题不在于解决一些复杂的统计学魔方难题，而在于它试图在一个只能捕捉阴影的洞穴墙上体现柏拉图式完美的公平形式。很难预测我们会采用哪种解决方案，以及当这些解决方案与监管和经济激励措施相互作用时，会产生什么样的成本。算法公平性，**从根本上来讲，是一个社会道德问题。**世界发展终究还是人的发展。科技发展仍要加大人文关怀，增加人文底蕴的支撑科 技才会不至于沦为技术的奴隶。从科技伦理的角度来看，未来实现科技与人文的统一才是科技发展的重中之重。 | |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |