当前人工智能伦理的现状及未来举措

一、研究背景

近年来，随着人工智能系统的广泛应用，大众对人工智能系统如何收集、分析和使用海量数据的担忧正与日俱增，系统收集的海量数据中可能包含许多个人隐私信息，并且在大众毫不知情的情况下被用于商业目的或其他未经准许的用途。人工智能系统应用在现实世界中还存在许多的潜在危害，例如人脸识别系统对于种族的歧视、简历审查系统对于性别的歧视等，这将会带来歧视、影响和扩大社会的分歧、产生错误信息。越来越多的人们意识到了人工智能系统的不当应用给社会带来的潜在危害，对人工智能伦理加强研究和监管的呼声变得越来越高。

在这种严峻的形势下，越来越多的公司、研究机构、政府机构和非盈利组织等参与到了人工智能伦理的研究中来，发布了许多的人工智能伦理原则和指南，在技术层面上对当前人工智能数据集和模型存在的偏见、公平性等问题进行了深入探讨研究，在产品层面上也进行了积极的应用实践。如今，人工智能伦理研究不再是纯粹的学术追求，已经成为了具有广泛社会影响的主流研究领域。

二、偏见和公平性

日前，斯坦福大学发布了《人工智能指数报告 2022》[1]，其中第三章全章论述了人工智能技术伦理。今年的 AI 指数关注被社区采用的指标在消除偏见和促进公平方面的进步。追踪这些指标在技术层面上的性能使得我们能够更加全面综合地看待公平性和偏见随着系统改善是如何改变的过程，随着人工智能系统越来越广泛的部署，达成这种理解是十分重要的。

在过去的几年里，我们投入了很大努力创建数据集、基准和指标来衡量机器学习模型的偏见和公平性。偏见是从训练数据中学习到的，训练数据中的偏见来自于社会的系统性偏见或者收集整理训练数据的人员。公平性表现在算法做出既不偏袒也不歧视独立的个人或群体的预测，此时算法被认为是公平的。人工智能系统在伦理方面的衡量主要采取基准和诊断指标这两种形式，基准是衡量整个领域进展的有用指标，诊断指标使研究者理解系统对特定应用或群体的潜在危害。文章中详细介绍了自然语言处理任务中的基准和指标在性别、种族、职业、残疾、宗教、年龄、外貌、性取向和种族等方面衡量偏见的效果。

报告中着重表述了三点：（1）语言模型比以前更有能力，但是新的数据同时表明越大的模型越有能力从训练数据中反映偏见。（2）人工智能伦理研究不断兴起，变得无处不在。文中分析了 ACM 会议在公平性、责任和透明性（FACCT）方面发表论文的领域、数量等，还分析了 NeurIPS 会议在可解释性、因果推理、隐私和数据收集、公平性和偏见等方面的论文，力证了人工智能伦理研究的活跃程度。（3）多模态模型学习多模态的偏差。多模态模型在语言-视觉任务上表现强大，但同时也反映了社会的刻板印象和输出中存在的偏见——文中详细分析了多模态模型 CLIP 上在性别、种族等方面所表现出的偏见。

三、人工智能伦理原则中的经验

为了消除公众对人工智能系统收集、使用和处理大数据的担忧、提倡用有

道德和负责任的态度开发人工智能系统，发布了许多人工智能伦理的原则和指南[2]。人工智能伦理原则对于建立共同理解和协同行动十分重要，是人工智能未来治理和创新的基础。通过研究发现，尽管人工智能伦理声称是全球性的并且对所有人都有益，但是它们在代表不同地区、实体和社区的价值观和观点时仍然存在局限性，资助人工智能伦理制定的实体和地区更多来自欧美国家。为此需要确保未来的合作、投资、标准、规范或立法具有能够反映各方声音的多样性和包容性。

此外，还要采取真正的行动，从高抽象的概念论证转向在实践中践行人工智能伦理，并且建立真正对公众利益有利的问责机制。在人工智能伦理原则的制定、践行中要始终记住“谁制定了人工智能管理的议程？该议程代表了什么文化逻辑，谁从中受益？”。从长远来看，人工智能伦理的制定与实践有助于人工智能行业的可持续和良性发展，更加尊重和服务社会需求而不是服从某个政府或私营公司的利益。

四、思考和行动

从客观上来看，社会是固有的、不可避免的存在偏见，并且这种偏见存在于社会的每一个个体中，偏见不可能被消除而只能削弱。如果一味地追求消除偏见、促进公平，会对机器学习模型带来性能上的下降，这在人工智能指数的研究报告中得到了印证，并且这种行为本质上也是引入偏见的一种过程。我们对于模型的偏见和公平性的追求不能太激烈，出现有违“伦理正确”的现象就对模型全盘否定，出现这种情况应该对模型进行干预调整而不是抹杀。

数据是真实的反映了偏见，偏见是客观存在于数据中或者收集数据的人中的。“社会职业和资源分配本就与种族、性别、肤色相关联而分工不同，这难道不是事实么？至少现在是事实。”我们所要做的是别让偏见扩大和影响了我们的社会价值观，尽可能地让模型符合绝大多数人的共同价值观和群体利益。在人工智能伦理原则的制定、实施上，要符合绝大多数人的利益、为公众代言，而不是某个政府或组织，要把人工智能伦理原则真正地用在产品服务的实践上，让抽象的概念论证转化为实践，而不是束之高阁，真正使这些人工智能伦理原则和指南促进人工智能产业的可持续和良性发展。

参考文献

[1] Zhang D， Maslej N， Barbe A , et al. The AI Index 2022 Annual Report[J]. 2022.

[2] Hickok M . Lessons learned from AI ethics principles for future actions[J]. AI and Ethics, 2020(7697).