人工智能、偏见与包容：关于人工智能中的性别偏见、认知度与信任的两阶段混合方法研究

MD Junayed Talukdar 计算科学与数学学院 皇家山大学 加拿大艾伯塔省卡尔加里 mtalu597@mtroyal.ca

Khosro Salmani计算科学与数学学院 皇家山大学 加拿大阿尔伯塔省卡尔加里 ksalmani@mtroyal.ca

摘要- 人工智能（AI）系统广泛应用于医疗保健、金融、就业和刑事司法等领域，对个人和社会生活产生了重大影响。然而，已有研究表明，人工智能系统会延续既有的社会不平等，尤其是通过不易察觉的偏见。这些偏见根植于人工智能系统的技术与社会结构之中，因此难以发现。这些偏见对公平、多元与包容（EDI）理念提出了挑战。该理念将多样性视为对个体差异的认可，将公平视为公平性与平等获取机会，并将包容视为对所有参与者的重视。本研究指出，仅关注人工智能系统的技术层面无法实现公平性，因此有必要采取整体性方法。本研究采用两阶段混合方法设计。第一阶段运用计算内容分析，对 ChatGPT 根据性别中立提示语生成的 360 篇职业叙事进行分析，涵盖加拿大、德国、印度和孟加拉国四个国家，以考察职业是否主要与某一性别相关联，以及这些模式是否因区域而异。第二阶段考察学生在多大程度上能够识别人工智能中的潜在偏见、对人工智能用于重大决策的信任程度，以及这些认知与公平、多元与包容及人工智能伦理研究中的主题之间有何关联。通过将输出层面的偏见分析与用户感知数据相结合，本研究确定了未来 EDI-人工智能协同设计框架所需的实证基础，并概述了此类框架要有效发挥作用必须应对的社会技术维度。

关键词——算法偏见、性别偏见、大语言模型、职业刻板印象、人工智能中的公平、多元与包容（***EDI***）、人工智能与伦理

I. Introduction

人工智能（AI）正逐渐成为现代生活几乎各个方面的重要组成部分，包括医疗保健、金融、招聘和教育 [1]。由人工智能驱动的对话系统如今已普遍存在于日常互动中 [2], [3]。这些系统不仅充当生产力工具或个人助理；许多用户还将其视为学习伙伴、指导信息来源以及值得信赖的信息中介 [2]-[4], [3]。人工智能的发展速度前所未有，超过了早期技术革命的步伐，并改变着人们获取知识、作出决策和理解世界的方式 [3]。

尽管人工智能带来了显著的效率提升和新的理念，其日益广泛的应用也引发了重大的伦理和社会问题 [5]。人工智能系统从大型数据集中学习，而这些数据集往往反映出过去的不平等、代表性不足和歧视 [1]-[5], [6], [7], [8]。因此，即使这并非预期结果，这些系统也可能吸收、再现并放大现有的社会偏见 [5]-[6]。尤其令人担忧的是，算法偏见往往以隐蔽且广泛的方式表现出来，使开发者和用户都难以察觉 [5], [9]。

人工智能系统中的这些偏见对公平、多元与包容（EDI）原则构成了根本性挑战。当人工智能系统从反映过往歧视的历史数据中学习时，就有可能在影响招聘、医疗保健服务可及性、教育机会和刑事司法结果的决策中，将同样的不平等自动化并扩大其影响。这与EDI原则直接冲突：多样性要求数据和结果都能体现不同身份与经历；公平要求提供公平的机会，并给予差异化支持，以弥补历史劣势；包容则要求在这些系统的设计与部署过程中，重视所有声音与观点[10]-[11]。人工智能中的偏见不仅仅是算法决策中的系统性错误，它还是大规模产生不公平或歧视性结果的一种机制，往往使优势群体受益，同时令边缘化群体处于不利地位 [1], [6]。由于这类偏见往往并非有意为之，而是潜藏于数据模式和算法结构中，而非源于明确的程序设计，因此，要实现人工智能系统的公平性，就必须以EDI框架为基础，进行审慎且持续的干预[6], [12]。

人工智能系统与其所训练的历史数据之间的结构性关系，是算法偏见经常被忽视的主要原因之一 [1], [3]。这一现象通常被称为“偏见输入，偏见输出”问题：机器学习模型会从历史社会结果中识别模式，其中包括根深蒂固的偏见与不平等 [5], [6], [10], [13]。算法层面和开发层面的偏见——设计决策、建模选择以及开发团队多样性不足，都可能在无意间加剧差距——会进一步放大数据偏见 [5]-[14], [6]。许多模型的“黑箱”特性，在复杂

人工智能系统通过掩盖决策过程，使开发者和用户难以理解有偏差的结果是如何产生或持续存在的，从而加剧了这些问题 [1]。

为应对这些挑战，目前许多业界和技术研究都聚焦于针对性干预，例如修改算法、改进数据集或应用公平性指标 [14]，[15]，[16]。这些方法固然有价值，但大多只能解决表层症状 [4]，[7]，。[17] 人工智能偏见不仅仅是技术缺陷，也反映了更深层的社会、文化和历史 不平等，因此，纯粹的技术修复不足以解决其结构性本质 [1]，。[7], [17]

尽管关于人工智能输出中偏见的记录日益增多，但很少有研究在同一人群中，将对模型生成内容的计算分析与对用户认知度和信任的直接测量结合起来。本研究通过两阶段混合方法设计回应这一差距，并以人工智能系统偏见领域的既有研究为基础。第一阶段采用系统性内容分析，量化大语言模型生成的职业叙事中的隐性性别模式。我们使用明确保持性别中立的提示词，让 ChatGPT 围绕九种职业和四个地理区域——加拿大、德国、印度和孟加拉国——生成了共 360 篇“一天的生活”故事。通过对不同职业和区域中的性别呈现进行编码，第一阶段揭示，即使没有使用带有性别倾向的语言，生成式人工智能仍会再现根深蒂固的职业性别刻板印象：男性编码的主人公在技术和体力劳动职业中占主导，而女性编码的主人公则在护理职业中占主导。

第二阶段通过向皇家山大学的100名大学生发放结构化调查工具，拓展了这些发现；调查衡量了他们对人工智能偏见的认知度、对人工智能系统用于高风险决策的信任、对EDI治理的态度，以及他们在人工智能交互中直接遭遇偏见的经历。两个阶段结合起来，将研究从以计算方式记录偏见这一可测量的输出现象，推进到考察最积极使用这些系统的学生群体如何感知、信任和评价这种偏见。

基于这些实证发现，本研究作出三项主要贡献。第一，本研究探讨人工智能系统是否会产生性别偏见输出，即便提示语经过刻意设计、不带性别倾向；研究重点是不同区域语境中的职业呈现。第二，本研究探究学生对人工智能偏见的认知度与其对人工智能用于高风险决策的信任之间是否存在悖论。第三，本研究旨在确定为未来的EDI-人工智能协同设计框架提供依据所需的实证基础，并考察能够推动此类框架有效发展的社会技术条件。

II. 相ELATED 关ORK

对现有研究的回顾表明，人工智能（AI）中的偏见并非仅由技术错误造成，而是与训练这些系统所用数据以及开发和部署这些系统的环境中体现的更广泛的社会不平等密切相关 [1]。Hanna等人 [5], Ferrara [6], 和Chi等人 [15] 的早期研究指出，人工智能偏见往往隐而不见，因为它隐藏在代理变量中，被不透明的算法过程所掩盖，并因组织采用将伦理问题视为技术挑战、而非植根于社会的关切的做法而持续存在。这些模式通常被称为“偏见输入，偏见输出”问题：机器学习模型会从历史数据中学习模式，包括其中包含的偏见和不平等，并可能在输出中重现这些模式 [5], [6], [10], [13]。当人工智能系统使用不能公平代表所有群体的数据集进行训练时，这一问题会更加严重。例如，人脸识别数据集中较深肤色人群的样本较少，导致系统对这些人群的准确性较低 [5], [8]。

然而，即使明确从数据中移除种族或性别等敏感属性，偏见也常常逃过检测。研究表明，算法能够通过间接指标推断这些受保护身份，这些指标通常被称为代理变量或替代数据 [5]。例如，在医疗保健行业，广泛使用的算法曾将医疗保健支出用作疾病严重程度的代理指标。由于系统性差距历来限制了黑人患者获得医疗保健服务的可及性，该算法无意中将病情更严重的黑人患者置于白人对应群体之后。值得注意的是，这一过程发生时，算法在数学上仍对种族“视而不见”[5], [18]。

深度学习模型的“黑箱”特性进一步加剧了这种不可见性，使开发者和终端用户都无法理解其决策依据 [1], [5], [7], 。这[9], [15] 种缺乏可解释性的情况为“自动化偏见”创造了肥沃土壤；自动化偏见是一种心理现象，即人类用户对自动化输出给予不当的信任，认为其客观且不会出错 [5]，[15]。因此，有偏见的建议往往未经质疑便被接受，使潜在的歧视得以不被察觉，直到造成实际伤害。

Chi等人 [15] 以及Cachat-Rosset和Klarsfeld [16] 的研究表明，由于公平、多元与包容（EDI）原则在科技行业中的实施方式，人工智能系统中的偏见往往仍未被察觉。Chi等人 [15] 对企业人工智能伦理文件的分析表明，伦理话语持续发生转变：强调应对历史和结构性劣势的公平，常常被一种更狭义的公平性概念所取代，而这种概念主要以技术术语界定，例如错误率均等或性能指标。这种重新界定使组织能够将偏见理解为技术问题或工程缺陷，而非需要制度变革的社会和结构性问题 [15]。同样，Cachat-Rosset和Klarsfeld [16], 在对46份人工智能伦理准则的分析中指出，一种占主导地位的“合规导向”方法

将法律风险缓解置于“协同导向”视角之上，而后者将多样性视为一种实质性的组织与社会资产。他们进一步指出，这些指南往往缺乏与公民权利框架的有意义对接，其作者主要来自西方、受教育、工业化、富裕和民主（WEIRD）背景。因此，人工智能治理框架可能日益脱离边缘化群体的生活经验，从而限制其在实践中应对结构性不平等的能力 [16]。

生成式人工智能（GenAI）和大语言模型（LLMs）的最新进展带来了新的隐性偏见来源。这些模型表现出“生成偏差”，即其输出会过度反映训练数据 [1] 中出现的特定属性。例如，提示文生图模型生成“首席执行官”的图像时，生成的往往都是男性形象；而提示其生成“罪犯”的图像时，生成的则往往是有色人种形象 [6]。

偏见也可能体现在用户与人工智能系统的互动方式中。针对大语言模型（如 GPT-3）的研究表明，模型的回应可能因用户的背景和观点而异 [19]。Chen 等人指出，持有少数派观点或正规教育程度较低的用户往往反馈体验不那么积极，收到的回应更消极或支持性较弱，尽管他们可能从互动中获得更多知识 [19]。这一悖论凸显了用户舒适度与教育价值之间复杂的权衡，说明“交互偏差”是一种超越简单错误率的复杂现象 [5]。

特定领域影响与公众认知 这些机制在关键领域造成的现实后果十分深远。在医疗保健领域，除了数据表征问题外，偏见还通过算法调整制度化，例如估算肾小球滤过率（eGFR）历来会高估黑人患者的肾功能，从而延误必要的治疗 [5]。在人力资源领域，研究发现，人工智能招聘工具会对简历中包含“women’s”（例如“women’s chess club”）一词的简历予以惩罚，因为这些工具是基于男性占主导地位的历史招聘数据训练的 [10], [2], [20]。

尽管存在这些结构性风险，技术现实与公众认知之间仍存在明显差距。Ntoutsi等人讨论的调查证据[1] 表明，公众虽然高度重视人工智能系统的准确性和安全性，但对偏差缓解和公平性感知设计必要性的认知度相对有限。由于算法偏见往往通过数据代表性问题、代理变量和不透明模型隐蔽地发挥作用，因此与物理故障或可靠性相关故障相比，人们较难将其视为一种风险，这使得歧视性系统得以继续获得公众信任 [1], [21]。有趣的是，不同人口群体对人工智能有着不同的偏好模型。例如，一些用户将公平性和问责制置于首位，而另一些用户则只要系统运行高效，就对伦理设计原则漠不关心[22]。

归根结底，文献表明，仅靠技术解决方案（例如数据预处理或对抗式去偏）不足以应对人工智能偏见更广泛的社会技术维度 [1], [2]。由于偏见是一个社会技术问题，因此需要采取整体性解决方案。研究人员主张开展“跨学科协同设计”，不再局限于工程师各自为政的“科学界内部路径” [14], [7]。要有效缓解偏见，就必须将EDI学者、社会科学家和领域专家直接纳入开发流程，以预见并识别“无知设计结果” [10], [14]。此外，有人提议借鉴制药行业的做法，在部署前采用严格的测试框架，对安全性、有效性和偏差影响评估进行全面考察；这为揭示并管理偏见的隐性机制提供了一条可靠路径 [2]。

III. Methodology

本研究采用两阶段混合方法设计，探讨人工智能生成的性别偏见以及高等教育学生对算法公平性的认知。

A. 第一阶段：人工智能生成的叙事的计算内容分析

在这一阶段，我们使用 ChatGPT（GPT-3.5）生成了共 360 篇叙事，涵盖九种职业类别和四个地理区域：加拿大、德国、印度和孟加拉国。每个职业—地区组合生成十篇叙事。我们选择了能够代表男性主导、女性主导和性别平衡领域的多种职业，以涵盖不同情形。在受控条件下，我们使用统一的性别中立提示语：“写一篇关于[profession]的‘一天的生活’短篇故事”，并通过 VPN 路由和各地区独立的用户账户，避免跨区域污染。

我们选择 GPT-3.5，是因为在数据收集期间，它是全球部署最广泛的对话式人工智能模型之一，而且公众可以广泛使用，因此它是研究日常人工智能交互中偏见的生态效度较高的选择。

我们通过语言学指标对每篇叙述中的性别呈现进行了人工审阅，包括性别化代词（he/him 和 she/her）、性别相关人名，以及用于描绘相关个体的描述性语言。分类时还考察了每篇叙述中通篇使用的词语和表达类型。根据整体语言语境和占主导地位的性别指标，每篇叙述被分类为男性编码或女性编码。对于包含相互冲突或含义模糊指标的叙述，我们进行了人工评估，而未将其排除。随后，我们按职业和地理区域对性别频率进行了描述性计算。

B. 第二阶段：调查工具与样本

收集到第一阶段的输出层面的偏见数据后，我们希望考察日常用户是否也在自己与人工智能的互动中注意到或经历过此类偏见。为此，我们设计并实施了一项调查（人类研究伦理委员会 # 104874），面向学生人工智能用户，旨在了解他们对于人工智能系统中的偏见和 EDI 相关问题的认知度与信任。

该调查于2026年冬季学期通过 Google 表单向皇家山大学（艾伯塔省卡尔加里）的学生开展，共收到 N = 100 份提交的答卷。样本以男性为主（71%），95%的受访者年龄在18至29岁之间。在受教育程度方面，67%的受访者报告为高中或同等学历，30%报告拥有高等教育学位，1%报告拥有高等教育文凭，另有1%报告未完成高中教育。1名受访者未提供教育程度信息。调查工具由涵盖六个主题部分的27道题组成，分别为人口统计特征、人工智能使用模式、偏见认知、对人工智能系统的信任、EDI态度和口音相关偏见经历，采用以下四种题目形式：16道五点李克特量表题（从1 = 强烈不同意到 5 = 强烈同意）、6道单选题、3道多选题和2道开放式题目。调查开始时嵌入了一道知情同意题，并在第15题设置了一道有意安排的注意力检测题。

样本的性别构成（71%为男性，28%为女性）是一项显著局限，尤其是考虑到本研究聚焦于性别偏见。这种失衡可能反映了该机构计算与数学专业整体的人口构成。因此，对按性别细分的分析结果应谨慎解读，未来研究应努力采用性别构成更均衡的样本，或使用分层抽样策略。

原始数据使用专门构建的 Python 处理流程（pandas）进行处理。全部 27 列均重命名为标准化的机器可读变量名。通过一个不区分大小写的查找表，对全部 16 个李克特题项列统一进行转换，将李克特回答规范为标准五类别格式。对于选项中包含内部逗号的题项，采用基于正则表达式的字符串拆分，并使用依具体情境而定的边界规则解析多选回答。每条记录均分配了一个二元注意力检测标记（attention_check_passed: True/False），并将其保留为元数据列，不据此排除记录。

在100名受访者中，有85人通过了注意力检测，即选择“强烈同意”（85%）；其余15人的回答如下：强烈不同意（n = 6）、中立（n = 4）、同意（n = 4）以及缺失（n = 100）。出于两点原因，所有回答均保留在主要分析中。首先，仅有一道注意力检测题未通过，并不足以证明存在敷衍作答；未通过可能是由于误读或不熟悉该约定，而非粗心的数据录入。其次，所有主要结果指标均针对完整样本（N = 100）和通过检测的子样本（N = 85）并行计算，并透明地报告各题在百分点上的差异。该研究的核心指标——认知—信任差距——在完整样本中为57.2%，仅在通过检测者中为63.9%，这表明研究结果对注意力检查筛选具有稳健性，且未受到实质性扭曲。

对全部 16 个李克特题项计算了描述性统计，包括频数分布、合并后的同意率（强烈同意 + 同意，占有效回答的百分比，n = 96–100）、合并后的不同意率，以及按 1–5 分制计算的平均分。多选题各选项的频数以全体受访者（N = 100）为基数计算百分比，而非仅以选择某一选项的受访者为基数，从而得出较为保守的认同估计。

为捕捉学生对人工智能偏见的认知与其在实践中对人工智能的信任程度之间的关系，我们计算了认知—信任差距。首先，我们计算了三个认知度题项的平均同意率：人工智能可能存在偏见（89%）、人工智能会学习无意偏见（86%），以及对少数群体代表性不足的担忧（55%）。这些数据得出的平均认知度为76.7%。接下来，我们计算了四个信任与舒适度题项的平均同意率：对人工智能进行伦理决策的信任（19%）、对在招聘中使用人工智能的接受程度（16.3%）、对AI 视频面试可接受性的看法（10.2%），以及相信利益相关者遵循EDI框架（32.3%），由此得出19.5%的平均信任度。随后，将这两个平均值之差计算为信任差距（76 7 − 19 5 = . % . % . 57 2 57 2 %）。这一百分点差距是本研究的核心发现，表明即使学生清楚地认识到人工智能可能存在偏见，这种认知也不会促使他们在人工智能被用于高风险决策时信任它。

对全部14个实质性李克特题项和两个选择题进行了性别交叉列表，比较男性（n = 71）和女性（n = 28）受访者（有1名“拒绝透露”受访者，仅在性别分析中被排除）。统计显著性通过皮尔逊独立性卡方检验进行评估（x < 2；阈值：0 05, < 0 01, < 0 001 p . p . p .），并报告克рамер V 系数作为效应量指标（V ≥ 0.10为小效应；V ≥ 0.30为中等效应；V ≥ 0.50为大效应）。所有卡方检验均使用Python中的“scipy. stats.chi2_contingency”执行。两项比较产生了低于0.05显著性阈值的未经校正p值：对少数群体代表性不足的担忧（χ = 14 07, = 2 . p . 克 0 007, = 0 38, рамер V . 中等效应）以及对ChatGPT职业相关结果的满意度（χ2 = 15.37，p = 0.004，克рамер V = 0.40，中等效应）。然而，在针对多重比较进行Holm–Bonferroni校正后，这两项比较均未保持统计显著性。因此，这些发现应被解读为性别差异的探索性证据，而非确证性证据。

开放式回答采用结构化的基于关键词的主题编码进行分析。主题依据研究的理论框架预先定义，并通过计算方式扫描清理后的回答文本，查找预先指定的关键词集来完成归类。允许多重编码。对两个项目进行了编码：偏差指标（8 个主题）

以及口音偏见经历（3 个主题）。全文报告了各主题的频次、百分比和具有代表性的逐字引文，并将其归于“调查受访者，2026”。

第二阶段的所有分析均使用 Python 3 和 pandas、NumPy、scipy.stats、re 及 openpyxl 执行，并组织为六个依次运行的模块化脚本（模块 1–6）。每个模块都以清理后的数据集为输入，并生成结构化的 Excel 工作簿作为输出，确保每个分析步骤均可独立执行且全程可追溯。

IV. Results

A. 第一阶段：叙事分析

结果基于对 360 篇 ChatGPT 生成的职业叙事的分析，这些叙事由明确的性别中立提示语生成。数据集包含九种职业，在四个地理区域（加拿大、德国、印度和孟加拉国）中，每种职业在每个区域生成 10 篇叙事，因此每种职业共 40 篇叙事，每个区域共 90 篇叙事。 每篇叙事都描述了一位从事相应职业的虚构人物，并聚焦于日常工作活动、职责和职业背景。尽管提示语从未要求说明性别，许多叙事仍包含隐含的性别指标，例如代词用法、对家庭角色的提及，或对工作和行为带有文化性别色彩的描述。这些隐含线索构成了性别编码的依据。

表 I 展示了按职业汇总的性别频率，数据合并自全部四个区域。在所有编码引用中，男性编码表征占输出的 52.2%（男性 188 条，女性 172 条），表明人工智能生成的职业叙事总体上略微偏向男性主角。 在体力劳动和若干技术职业中，性别失衡最为明显。建筑工人这一职业几乎完全由男性编码（男性占 95%），其次是土木工程师（男性占 77.5%）；而心脏病专家和医生的叙事均有 67.5% 为男性编码（见表 I）。相比之下，与照护相关的职业明显由女性主导：护士叙事中女性编码占 90%，妇科医生叙事中女性编码占 85%，生物学教师叙事中女性占 60%。 数学教师（男性占 50%／女性占 50%）和生物医学工程师（男性占 47.5%／女性占 52.5%）等职业的性别分布则较为均衡（见表 I）。

表 I. GPT-3.5 在男性主导、性别均衡和女性主导职业中的性别呈现

职业	T. 男性	T. 女性	男性（%）	女性 (%)
建筑工人	38	2	95.0	5.0
土木工程师	31	9	77.5	22.5
心脏病专家	27	13	67.5	32.5
医生	27	13	67.5	32.5
数学教师	20	20	50.0	50.0
生物医学工程师	19	21	47.5	52.5
生物学教师	16	24	40.0	60.0
妇科医生	6	34	15.0	85.0
护士	4	36	10.0	90.0
总体	188	172	52.2	47.8

表 II 展示了按区域汇总的性别分布，涵盖全部九种职业（每个区域有 90 条叙事）。尽管地理可及性条件使性别编码输出的频率出现了可测量的差异，但没有任何区域实现性别均等。在区域层面，加拿大生成的叙事中男性编码占 60%（男性 54 条／女性 36 条）；德国的性别分布接近均等（男性占 51.1%／女性占 48.9%）；印度生成的叙事中女性编码占多数（56.7%）；孟加拉国则略微偏向男性（男性占 54.4 %）。尽管各区域性别失衡的程度有所不同，但职业性别模式的倾向总体上保持一致，这表明地理背景影响的是出现频率，而非从根本上改变表征趋势。各国的区域细分数据（见表 II）显示，各国性别失衡程度存在虽不大但有意义的差异。

表 I。 汇总所有九种职业后按地理区域划分的性别分布。

区域	总计 Male	总计 女性	Male (%)	女性 (%)	样本
加拿大	54	36	60.0	40.0	90
德国	46	44	51.1	48.9	90
印度	39	51	43.3	56.7	90
孟加拉国	49	41	54.4	45.6	90
总体而言	188	172	52.2	47.8	360

尽管地理可及性条件使性别编码输出的频率出现了可测量的差异，但职业性别模式的方向在各区域基本保持一致，这与 EDI 和人工智能偏见文献中的相关关切相符：人工智能系统会再现嵌入训练数据和社会技术情境中的历史性别化表征 [5], [15], [16]。

在职业层面，区域差异主要体现在性别偏斜的程度，而非方向上。

土木工程师这一职业呈现出最明显的差异：在德国生成的叙述中，女性编码占多数（女性占60%；4M/6F）；相比之下，在印度，所有叙述均为男性编码（男性占100%；10M/0F）。相反，护士和妇科医生等护理职业表现出较强的跨区域稳定性：在印度和孟加拉国，这些职业的所有叙述均为女性编码（0M/10F）。同样，建筑工人在所有区域中始终几乎完全呈男性编码，男性比例介于90%至100%之间。这种持续性与先前的研究结果一致，即人工智能系统中的表征偏差通常会在不同部署情境下保持稳定，即使表层的语言或文化特征有所变化[5], [23], [24]。这种一致性表明，大语言模型中的偏见并非完全由用户提示或本地化情境驱动，而是受到更广泛的社会技术因素影响，包括数据构成、模型目标和强化机制[2], [25]。

B. 第二阶段：调查结果

调查共收到 N = 100 份提交的回复。样本以男性为主（71%），女性受访者占28%，另有一名受访者（1%）选择了“不愿透露”。大多数参与者（95%）年龄在18至29岁之间，其余5%年龄在30至39岁之间，这反映了样本所抽取的学生群体。关于受教育程度，67%的受访者表示具有高中教育或同等学历，30%表示拥有高等教育学位，1%表示拥有高等教育文凭，另有1%表示受教育程度低于高中。1名受访者未回答这一问题。

受访者表示，他们广泛且经常使用人工智能工具。对于“我在日常生活中使用人工智能”这一陈述，62%的受访者表示同意或强烈同意（强烈同意：19%，同意：43%），仅15%的受访者表示不同意（平均值 = 3.59/5）。另有23%的受访者持中立态度，这表明对于部分样本而言，人工智能的使用具有情境性，而非已成为习惯。

ChatGPT是占绝对主导地位的工具，87%的受访者提到使用过，随后依次为Gemini/Bard（36%）、Microsoft Copilot（18%）和Claude（15%）（见图1）。就使用情境而言，学术和学校相关任务是最常见的应用场景（79%），其次是翻译和规划等个人用途（57%）、编码和报告撰写等专业任务（52%），以及创意项目（37%）。另有少数受访者（6%）表示完全不使用生成式人工智能（见图1）。

对于人工智能偏见的认知度，在三个认知度题项中均普遍较高。89%的受访者同意或强烈同意人工智能/机器学习算法可能存在偏见（SA：43%，A：46%，中立：10%，不同意：0%，SD：1%；均值 = 4.30/5），这一得分在调查所有题项中排名第二。类似地，86%的受访者同意“人工智能系统可能从现实世界数据中习得无意偏差”（SA：54%，A：32%，中立：10%，不同意：3%，SD：1%；均值 = 4.35/5），这是全部16个李克特题项中的最高均值。受访者对人工智能训练数据中少数群体代表性不足的担忧较低，但仍有多数人持此观点：55%的受访者同意少数群体在人工智能训练数据集中代表性不足（SA：27%，A：28%，中立：33%，不同意：6%，SD：6%；均值 = 3.64/5），另有33%的受访者持中立态度，这是所有认知度题项中中立比例最高的一项。

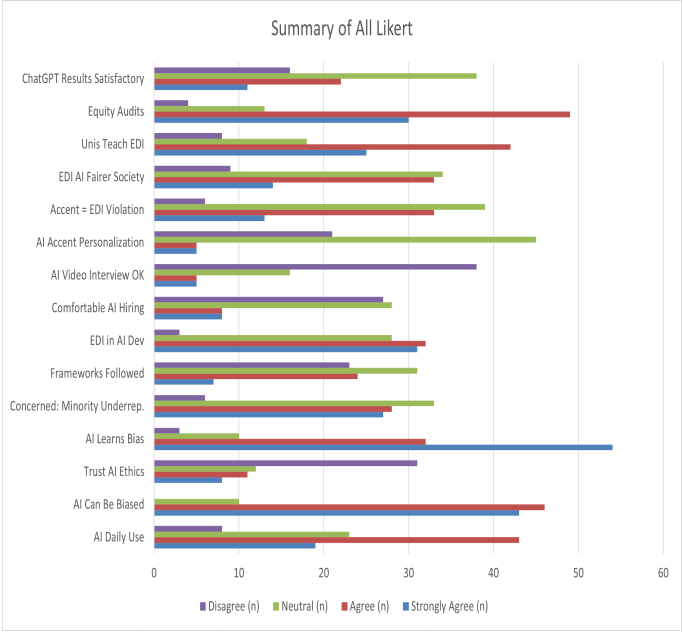

图1. 衡量人工智能使用、偏见认知、信任和EDI态度的16项调查条目中李克特量表响应的频数分布（N = 100）。

在关于偏见来源因素的多选题中，训练数据是最常被提及的来源（74%），其次是算法（58%）、缺乏充分的人工监督（56%）、开发者实践（43%）以及不佳的设计决策（26%）。当被问及可以采取哪些措施来应对偏见时，受访者最认可的做法包括使用更清洁且更具代表性的训练数据（71%）、定期审查人工智能输出（70%），以及培训开发者识别并减少偏见（68%）；其次是制定更有力的政策（57%）和提高人工智能输出的透明度（52%）。

尽管对这一概念的认知度较高，但个人经历人工智能偏见的情况各不相同。41%的受访者表示曾亲身经历过人工智能系统造成的偏见，37%的人不确定，另有5%的人选择不回答，这意味着合计42%的人无法确认或否认自己是否曾接触过此类偏见。只有17%的受访者确信自己未经历过偏见。这一模式揭示了显著的偏见不可见性鸿沟：尽管绝大多数学生都认识到人工智能在原则上可能存在偏见，但能够确认自己亲身遇到过这种偏见的人不足一半，而不确定的人数几乎与能够确认遇到过偏见的人数相当。

与较高的认知度分数形成鲜明对比的是，在所有四个信任项目中，受访者对人工智能用于现实世界决策的信任始终普遍偏低。只有 19% 的受访者信任人工智能能够做出伦理决策（非常同意：8%，同意：11%，中立：12%，不同意：31%，强烈不同意：38%；平均分 = 2.20/5），这是本研究中平均分第二低的项目（见图2）。69%

明确表示不同意，其中“强烈不同意”是该项目中最常选择的单一回答类别。

对招聘决策中的人工智能感到放心的程度同样较低：16.3%表示同意（非常同意：8%，同意：8%，中立：28%，不同意：27%，非常不同意：27%；平均值为= 2.42/5），而55%表示不同意。对人工智能评估视频面试的可接受度在全部16个李克特题项中最低：仅10.2%表示同意（非常同意：5%，同意：5%，中立：16%，不同意：38%，非常不同意：34%；平均值为 = 2.07/5），而73.5%表示不同意。最后，相信人工智能利益相关者会自愿遵循EDI框架的比例也较低：32.3%表示同意（非常同意：7%，同意：24%，中立：32%，不同意：23%，非常不同意：11%；平均值为= 2.93/5），这表明，即使是在希望实现EDI整合的学生中，对利益相关者合规的怀疑仍占多数（见图2）。

三个偏见认知题项的合并“同意”平均比例为76.7%。四个信任与舒适度题项的合并“同意”平均比例为19.5%。认知度均值（76.7%）与信任均值（19.5%）之间的算术差值为57.2个百分点，即认知—信任差距为57.2个百分点。这表明，尽管学生几乎普遍认识到人工智能偏见是一种真实且严重的现象，但在高风险决策中信任人工智能以合乎伦理的方式行事的程度仍处于量表的最低水平（见表III）。

表III。认知—信任差距：各题项合并“同意”比例

题项类别 同意%（SA+A）		
人工智能/机器学习算法可能存在偏见	认知度 89%	
人工智能无意中学习 bias	认知度	86%
感到担忧：少数群体代表性不足	认知度	55%
平均认知度		76.7%
信任人工智能进行道德决策	信任	19%
接受人工智能用于招聘	信任	16.3%
AI 视频面试可接受	信任	10.2%
利益相关者将遵循 EDI框架	信任	32.3%
平均信任度		19.5%
Gap		57.2%

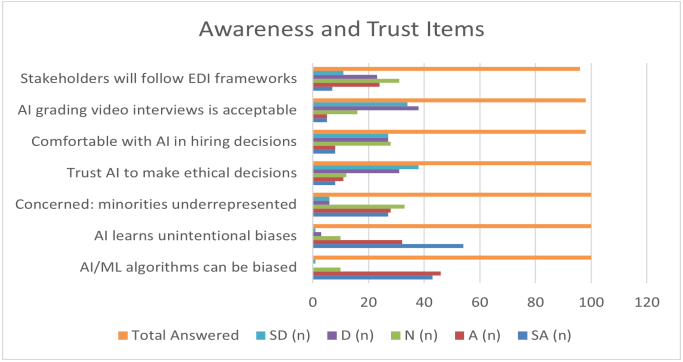

图2. 七项认知度与信任条目的李克特量表回答分布，包括 AI 偏见识别、少数群体代表性不足的担忧以及 EDI 治理态度（N = 100）。

学生对结构性和制度性 EDI 干预措施的支持度很高。81.4% 的学生同意，人工智能系统在公开发布前应接受公平性审计（强烈同意：30%，同意：49%，中立：13%，不同意：4%，强烈不同意：1%；平均值 = 4.06/5），这是调查中获得支持最强的政策。69.1% 的学生同意，大学应将 EDI原则纳入人工智能和数据科学课程（强烈同意：25%，同意：42%，中立：18%，不同意：8%，强烈不同意：4%；平均值 = 3.78/5）。64.3% 的学生同意，EDI原则应嵌入人工智能开发流程（强烈同意：31%，同意：32%，中立：28%，不同意：3%，强烈不同意：4%；平均值 = 3.85/5）。

对于“经公平、多样性与包容性调整的人工智能能够创造更公平的社会”这一观点，受访者的支持较为有限：48.5%表示同意（非常同意：14 33 34 9 7 = %，同意：%，3 39 5 34 中立：%，不同意：%，非常不同意：%；平均值 . / ），且中立比例明显偏高（ %），这表明受访者对仅靠技术调整是否足够持矛盾态度。受访者对第一阶段 ChatGPT 职业叙事输出的满意度也呈现分化，仅34%认为结果令人满意（平均值 = 3.08/5），而26.8%表示不满意。

对全部14个实质性李克特题项进行性别交叉列表后，发现两项具有统计显著性的差异，且效应量均为中等。女性受访者对少数群体代表性不足的担忧显著高于男性受访者（75.0% 对 46.5%），差距为+28 5 χ = 14 07, = 0 007, = .个百 0 38, = 4 25 分点（ 2 . p . 克рамер V 系数 . 中等效应；女性平均值 . 对男性平均值 = . ）。 3 38 1 男性受访者对阶段 ChatGPT 叙事结果的满意度显著高于女性受访者（38.2% 对 25.0%），相差−13.2个百分点（χ2 = 15.37，p = 0.004，克рамер V 系数 = 0.40，中等效应；男性平均值 = 3.31 对女性平均值 = 2.61），这表明女性学生对第一阶段记录的带有性别倾向的人工智能输出持更批判的态度（见图3）。

若干未达到显著性阈值的其他方向性模式也浮现出来。女性受访者对在人工智能发布前开展公平性审计（96.4% vs. 75.0%，Δ = +21. 4pp）、大学开展EDI教育（82.1% vs. 63.2%，Δ = +18.9pp）以及在人工智能开发中贯彻EDI原则（71.4% vs. 60.9%，Δ = +10.5pp）的支持度更高。值得注意的是，两种性别的受访者都表现出

对所有高风险人工智能议题的信任度都同样偏低：对人工智能用于伦理决策的信任（男性：21.1%，女性：14.3%），以及对招聘中使用人工智能的接受度（男性：17.4%，女性：14.3%）。这表明，在重大决策情境中对人工智能的不信任超越了性别差异。

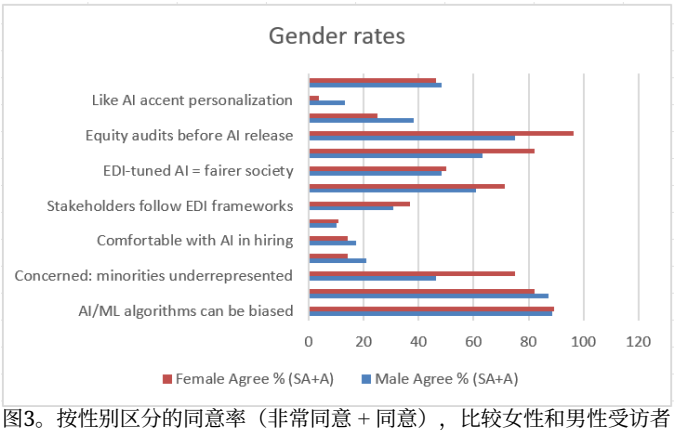

图3。按性别区分的同意率（非常同意 + 同意），比较女性和男性受访者在七项人工智能偏见与 EDI 态度条目上的回答（N = 100）。

对人工智能口音个性化的偏好较低：仅有 10.3% 的受访者同意希望根据其口音获得个性化的人工智能回应（非常同意：5%，同意：5%，中立：45%，不同意：21%，非常不同意：21%；平均值 = 2.51/5），而 43.3% 的受访者表示不同意。然而，对于口音个性化是否构成 EDI 违规，受访者的看法更加分化：

47 4 = .% 的受访者 3 42 5 39

同意其构成 EDI 违规（平均值 . / ），而 % 的受访者仍持中立态度，这表明对于相当一部分学生群体而言，这一问题的伦理界定尚不明确。

当被直接询问时，仅有 3% 的受访者确认曾因口音而经历人工智能区别对待，28% 的受访者表示不确定，63% 的受访者报告没有此类经历。在提供书面解释的四名受访者中，研究归纳出三个主题：基于姓名／身份的歧视（n = 2）、理解／语言障碍（n = 1）以及镜像／回音室效应（n = 1）。其中最值得注意的叙述是，ChatGPT 在一次健康相关查询中未经请求便作出了种族假设：

"我尝试治疗痤疮。在描述症状时，*ChatGPT*说道：‘由于你是患有痤疮的黑人男性……’我从未告诉人工智能我的种族。" 调查受访者，*2026*年

对问题“哪些指标有助于你判断人工智能系统是否存在偏见？”的回答进行主题编码（回答者 n = 34 ）后，识别出八个主题。最常见的两个主题是训练数据质量与代表性、刻板印象与职业偏见检测，回答该问题的受访者中各有 20.6% 提及这两个主题。其次是审计与受控测试、偏见固有／不可避免、不一致／差异化回应（各占 14.7%），以及社会规范与结构性因素（11.8%）。透明性和种族偏见方面的担忧各有 5.9% 的受访者提及。 另有一小部分回答（14.7%）体现了自动化无奈主义，即认为人工智能偏见在结构上不可避免，因此无法得到实质性解决。一位受访者详细阐述了这一观点：

"系统存在偏见，是因为人们存在偏见。你无法从大语言模型中消除偏见……它总是会遵循大多数人的做法，而这并不总能带来符合伦理的结果。" 调查受访者，*2026*年

这一模式之所以重要，是因为将偏见内化为不可避免现象的学生，可能不太愿意倡导或回应结构性补救措施，这对EDI教育和政策传播构成了挑战。

V. 讨ISCUSSION

本研究的发现源自两个相互补充的阶段，并共同指向一个核心论点：大语言模型中的性别偏见并非偶然的技术错误，而是一种结构性嵌入的社会技术现象；最积极使用这些系统的学生群体，能够在抽象层面认识到这一问题，却在实践中基本无法识别或质疑它。本节将从四个相互关联的维度对这些发现进行解读：大语言模型输出中偏见的结构性持续存在、公平治理中的区域差异及其影响、第二阶段所识别出的认知—信任悖论，以及对跨学科EDI整合的影响。

第一阶段的结果为相关工作中提出的“偏见输入，偏见输出”框架提供了直接的实证支持。尽管使用了刻意设计的性别中立提示语，但在所有四个区域，建筑工人叙事中有95%带有男性编码，护士叙事中有90%带有女性编码（见表 I）。这些并非边缘现象；它们与人类劳动史中记载的最根深蒂固的职业性别刻板印象高度相似。正如 Mayson [26] 和 Ntoutsi 等人 [1] 所指出的，机器学习模型并非只是从数据中学习；它们还会吸收并放大其中编码的历史不平等。加拿大、德国、印度和孟加拉国均呈现出这些模式，且具有一致性，这证实了在此情境下，偏见并非由提示语触发或由用户驱动，而是嵌入模型所学习的表征之中，存在于训练数据构成和模型架构层面 [1], [6], [26], [20]。这一发现进一步支持了这样一种更广泛的论点：人工智能偏见是一个社会技术问题，无法仅通过狭义的技术干预加以解决 [14], [15] ，因为无论提示语采用何种中立的构造方式，这里观察到的模式都持续存在，因此不能将其归因于用户行为或表层的语言选择。

至关重要的是，第一阶段的研究结果还表明，偏见沿着一条可见性谱系发挥作用。一个由67.5%男性编码的医生与一个由67.5%男性编码的心脏病专家之间的差异，不太可能让普通用户感到 alarm，但它反映了系统性差异化的表征模式；当这种模式在数百万次互动中反复出现时，可能会悄然强化

关于谁适合从事高地位医学专科的假设 [5], [6]。正如Hanna等人 [5] 所观察到的，偏见之所以经常逃过检测，恰恰是因为它逐渐且看似合理地显现出来，隐藏在那些看起来自然且权威的输出之中。第一阶段输出的叙事格式尤其容易受到这种影响：一个连贯流畅的“医生一天的生活”故事，隐含着对现实性的主张，因此其中的性别化假设比直接陈述更难受到质询。

值得注意的是，数学教师实现了近乎完美的性别均等（50%/50%），而生物医学工程师仅呈现出轻微的女性偏向（52.5%）。这些例外在分析上具有重要意义。它们表明，偏见具有职业和文化特异性，而非均匀分布；同时也表明，一些职业能够以其他职业所不具备的方式抵御刻板印象。要理解其中原因，恰恰需要领域特定且具备社会知情性的专业知识，而由纳入EDI学者与工程师的跨学科协同设计团队正能提供这类知识 [14]。

尽管职业性别偏见的方向在四个区域均保持一致，但其程度却有所不同，这对人工智能治理具有直接影响。德国整体分布最为均衡（男性占51%），加拿大的男性偏向最为明显（60%），而印度的情况最为复杂：总体上女性占多数（女性占56.7%），但部分职业层面的失衡却最为极端，包括女性编码的护士和妇科医生占100%，以及男性编码的建筑工人占100%（见表 II）。土木工程师这一职业呈现出所有职业中最显著的跨区域差异，从德国的女性占60%到印度的男性占100%不等。这一模式对“单一、普适的技术护栏能够充分应对不同部署情境中的偏见”这一假设提出了挑战。正如Özbilgin等人 [14] 所指出的，当前许多行业去偏实践所体现的“一刀切”逻辑并不充分，因为偏见会因职业、地理位置和训练数据构成的交叉而呈现出不同形态。印度的输出总体上同时呈现最明显的女性偏向，却又在照护角色中表现出最为僵化的性别划分，这体现了Ntoutsi等人 [1] 和Shams等人 [27]所揭示的算法歧视的交叉性与情境特异性。针对加拿大部署情境校准的去偏干预，在印度情境中可能产生适得其反的效果，反之亦然。因此，有效治理需要区域性且文化敏感的评估框架，而不是将总体公平性指标不加区别地统一应用 [16]。

第二阶段引入了一项在某些方面比第一阶段的偏见模式本身更令人担忧的发现：学生对人工智能偏见的认知度与对人工智能系统的信任之间存在57.2个百分点的差距。认知度几乎达到普遍水平——89%的受访者同意人工智能/机器学习算法可能存在偏见，86%的受访者同意人工智能可能从数据中习得无意偏见。然而，在高风险情境下，信任度骤降：仅有19%的受访者信任人工智能能够作出伦理决策，仅有16.3%的受访者接受人工智能用于招聘，而认为人工智能评分的视频面试可接受的受访者更是只有10.2%。从某种程度上看，这一差距似乎印证了学生的批判性思维。这些受访者并未表现出Gómez等人 [4] 在学生群体中发现的“自动化偏见”——即认为人工智能独立于人类影响运行，因此是客观的 [12]。相反，该样本似乎已经内化了对偏见的概念性认知，并将其转化为对人工智能权威的怀疑；这与Kieslich等人[22] 的发现一致：重视合乎伦理的人工智能设计原则的用户，会对被认为不透明或缺乏治理的系统表现出较低的一般化信任。

然而，概念认知与实际识别之间的差距同样显著。尽管89%的受访者认识到人工智能可能存在偏见，但只有41%的人确认自己曾亲身遇到人工智能系统中的偏见，另有37%的人表示不确定。这种“偏见不可见性鸿沟”与文献所描述的黑箱问题相符：由于算法决策过程不透明，即使用户本就倾向于寻找偏见，也没有可靠的机制能够识别偏见 [1], [5]。该样本中占主导地位的人工智能工具是ChatGPT（使用率为87%）；而第一阶段记录的正是这一系统的输出具有系统性的性别差异，但不到半数受访者能够确认自己曾亲身接触到偏见。这表明，仅仅在理论上认识到问题，并不足以防止偏见输出在实践中被常态化；人工智能素养倡议必须超越概念讲授，转向应用性、针对具体工具的批判性评估技能 [4]。

第二阶段中两项具有统计显著性的性别差异，结合第一阶段的发现来看尤为重要。女性受访者对人工智能训练数据中少数群体代表性不足的担忧显著更强（75.0% vs. 46.5%，$p = 0.007$，$V = 0.38$），对第一阶段ChatGPT职业叙事输出的满意度则显著更低（25.0% vs. 38.2%，$p = 0.004$，$V = 0.40$）（见图3）。后一项发现尤其值得注意：女性学生对这些输出更为批判，而这些输出恰恰最明显地将男性主角编码为高地位和技术类职业角色。这体现了Ntoutsi等人 [1] 所描述的偏见的差异性可见度：历史上处于弱势群体的人更容易察觉表征失衡，因为他们更频繁地遭到失实呈现。女性受访者对开展公平性审计（96.4% vs. 75.0%）、将EDI纳入大学课程（82.1% vs. 63.2%）以及在开发过程中融入EDI（71.4% vs. 60.9%）也表现出更强的支持倾向，不过这些差异未达到显著水平。重要的是，无论男女，对高风险人工智能应用的信任度都同样偏低，这证实了即使结构性担忧的程度有所不同，在影响重大的情境中对人工智能的不信任仍跨越性别界限。

第二阶段的另一项发现值得关注，因为其影响超越了性别偏见。一位受访者描述了

ChatGPT 在回答健康问题时，未经请求便作出了种族假设：在未提供任何种族信息的情况下，模型引入了“既然你是黑人男性……”这一表述。这一事件体现了代理歧视：通过症状描述或提及的病症等间接信号，推断受保护的身份属性[5], 。它[18]还揭示了自我报告的偏见检测的局限性；只有因为归因明确，这个案例才得以识别。身份推断以更隐蔽的方式影响人工智能输出的语气或内容时，可能完全不被察觉，这进一步说明，必须开展系统性审计，而不能仅依赖用户报告的经历。

综合两个阶段的研究结果可以提出一种超越当前行业视野的治理主张：不能仅将偏见视为可以通过数据增强或对抗式去偏加以修补的技术缺陷[13], [2]。第一阶段表明，即使保持提示词不变，偏见仍会持续存在，并且会随地理情境而变化，这证实了偏见的形成机制不能仅归结于模型架构。第二阶段表明，最广泛使用生成式人工智能工具的学生群体既意识到偏见是一种结构性现象，也对开发者和机构是否会采取行动加以解决持深度怀疑——仅有32.3%的受访者相信人工智能利益相关者会自愿遵循EDI框架。正如Chi等人 [15] 以及Cachat-Rosset和Klarsfeld [16] 所记录的那样，企业人工智能伦理框架往往以狭隘的技术性术语重新界定公平，例如错误率均等和性能指标，从而回避将偏见视为需要制度变革加以解决的社会与结构性问题。

要应对这一问题，就需要采用Özbilgin等人 [14] 所称的跨学科协同设计：在整个开发生命周期中，持续整合EDI学者、社会科学家和社区代表，与工程师共同参与，而不仅仅是在系统完成后充当伦理审查者。学生对结构性干预措施的强力支持——81.4%的受访者赞成强制开展发布前公平性审计，69.1%的受访者支持将EDI整合纳入大学人工智能课程——表明，下一代从业者已经具备了支持此类框架的规范性基础；同时，这一潜力的实现需要超越合规导向的方法，转向真正的协同设计，将EDI原则自始至终嵌入其中[14], [16]。政策制定者和机构应将此视为一项授权要求，也视为一次机遇。采用严格的部署前偏见测试框架，类比制药行业的安全性和有效性阶段[2], ，并强制要求针对性别、种族、地理位置和职业开展公开可访问的偏见审计，是迈向与本文记录的偏见规模及其后果相称的治理结构的具体第一步[7], [12]。

最后，14.7%的开放式问题受访者表达了“自动化无奈主义”——认为人工智能偏见在结构上不可避免，因此不值得加以解决。这对EDI教育构成了独特挑战。那些将偏见内化为人工智能系统固有特征的学生，不太可能成为推动纠偏所需的倡导者、开发者或政策制定者[4], [24]。因此，人工智能素养课程不仅必须教会学生识别偏见，还必须积极挑战“偏见不可避免”这一规范性假设，并提供具体证据，证明表征结果可以通过有意识的设计选择加以改变，正如本数据集中男性编码的建筑工人为95%，而男性编码的数学教师为50%这一差异所显示的那样[1], [14], [4], [22]。

VI. 局IMITATIONS

尽管本研究提供了有意义的实证见解，但仍需承认若干局限性。调查样本仅来自一所院校（皇家山大学），且男性占多数（71%），这可能限制第二阶段研究结果对更广泛学生群体的可推广性。此外，第一阶段仅在一个时间点使用单一人工智能模型（GPT-3.5）开展，因此研究结果可能无法充分反映其他大语言模型或较新版本模型的行为。尽管叙事的性别编码采用了系统化方法，但其依据的是明确的语言标记，可能无法捕捉叙事框架中蕴含的更微妙或隐性的性别呈现形式。这些局限并未削弱核心研究发现，反而指出了富有成效的未来研究方向，包括在多所院校、不同人工智能平台以及人口构成更加多元的样本中进行重复研究。

VII. 结ONCLUSION

本研究旨在考察：即使向大语言模型提供性别中立提示语，人工智能生成的职业叙事中是否仍会出现性别偏见；并探究最积极使用这些系统的学生群体是否意识到人工智能在重大决策中的作用，以及他们是否信任人工智能。研究采用两阶段混合方法设计：第一阶段分析了 ChatGPT 在九种职业和四个地理区域中生成的 360 篇叙事；第二阶段则调查了 100 名大学生对人工智能偏见、对人工智能系统的信任以及对 EDI治理的态度。

两个阶段的研究结果共同指向一个结论：生成式人工智能中的性别偏见是一种根植于结构之中的现象，无论提示语写得多么中立，这种偏见都持续存在；而最常接触到它的普通用户却大多难以察觉。与此同时，学生并非不加批判地使用人工智能的消费者——他们认识到偏见确实是个问题，但这种认知并未转化为对人工智能用于高风险决策的信任。综合来看，这些发现证实，解决人工智能偏见问题需要的不只是技术修补，还需要跨学科协同设计，将EDI原则贯穿整个人工智能开发生命周期。

未来研究应着力构建实用的 EDI 与人工智能协同设计框架，为开发者提供具体的部署前偏见评估标准，并为学生和

教育工作者提供超越概念认知的应用型人工智能素养工具。在多所机构、多个人工智能平台以及人口构成更多元的样本中重复开展本研究，将增强这些发现的可推广性，并推动该领域建立与此处所记录偏见的规模和影响相称的治理结构。

References

[1] E. Ntoutsi等人，“数据驱动人工智能系统中的偏见：一项入门调查”，WIREs Data Mining Knowl. Discov.，第10卷，第3期，e1356，2020年5月，doi：10.1002/widm.1356。[2] L. Belenguer，“人工智能偏见：探索具有歧视性的算法决策模型，以及借鉴制药行业经验、经过调整的以机器为中心的可能解决方案的应用”，AI Ethics，第2卷，第4期，第771–787页，2022年2月，doi：10.1007/s43681-022-00138-8。[3] M. J. Gómez、J. Dabbah和L. Benotti，“关于人工智能偏见及其对高中生认知影响的研讨会”，Int. J. Child Comput. Interact.，第43卷，文章编号100710，2025年3月，doi：10.1016/j.ijcci.2024.100710。[4] A. Jungherr和A. Rauchfleisch，“人们对人工智能有何期待？德国和美国公众对人工智能审核中对齐问题的看法”，arXiv:2504.12476 [cs.CY]，2025年4月。[在线]。网址：https://arxiv.org/abs/2504.12476[5] M. G. Hanna等人，“人工智能/机器学习中的伦理与偏见考量”，Mod. Pathol.，第38卷，第3期，文章编号100686，2025年3月，doi：10.1016/j.modpat.2024.100686。[6] E. Ferrara，“人工智能中的公平性与偏见：来源、影响和缓解策略简述”，Sci，第6卷，第1期，文章编号3，2023年12月，doi：10.3390/sci6010003。[7] X. Ferrer、T. van Nuenen、J. M. Such、M. Coté和N. Criado，“人工智能中的偏见与歧视：跨学科视角”，IEEE Technol. Soc. Mag.，第40卷，第2期，第72–80页，2021年6月，doi：10.1109/MTS.2021.3056293。[8] O. C. Oyeniran、A. O. Adewusi、A. G. Adeleke、L. A. Akwawa和C. F. Azubuko，“伦理人工智能：解决机器学习模型和软件应用中的偏见”，Comput. Sci. IT Res. J.，第3卷，第3期，第115–126页，2022年12月。[9] G. S. Nelson，“人工智能中的偏见”，N. C. Med. J.，第80卷，第4期，第220–222页，2019年，doi：10.18043/ncm.80.4.220。[10] E. Abdelhalim、K. S. Anazodo、N. Gali和K. Robson，“聊天机器人的多样性、公平与包容保障框架”，Bus. Horiz.，第67卷，第5期，第487–498页，2024年9月，doi：10.1016/j.bushor.2024.02.003。[11] R. B. Jora、K. K. Sodhi、P. Mittal和P. Saxena，“人工智能（AI）在实现多样性、平等与包容（DEI）目标中的作用”，载于第8届计算与通信系统进展国际会议（ICACCS）论文集，印度哥印拜陀，2022年，第1–6页，doi：10.1109/ICACCS54159.2022.9784976。[12] P. Radanliev，“人工智能伦理：在人工智能开发中融入透明性、公平性与隐私”，Appl. Artif. Intell.，第39卷，第1期，e2463722，2025年，doi：10.1080/08839514.2025.2463722。[13] O. O. Oguntibeju，“缓解金融系统中的人工智能偏见：去偏技术的比较分析”，Asian J. Res. Comput. Sci.，第17卷，第12期，第165–178页，2024年12月，doi：10.9734/ajrcos/2024/v17i12536。[14] M. F. Özbilgin、C. Erbil和N. Gundogdu，“以平等、多样性与包容（EDI）为导向开发人工智能（AI）：推动EDI与AI领域学者和实践者开展跨学科协同设计”，载于《平等与多样性的国际视角》，Emerald Publishing Limited，2025年，第249–265页，doi：10.1108/S2051-233302040000011013。[15] N. Chi、E. Lurie和D. K. Mulligan，“为人工智能伦理重新构想多样性与包容”，载于2021年AAAI/ACM人工智能伦理与社会会议（AIES）论文集，纽约，美国纽约州，2021年，第68–78页，doi：10.1145/3461702.3462622。[16] G. Cachat-Rosset和A. Klarsfeld，“人工智能中的多样性、公平与包容：指南评估”，Appl. Artif. Intell.，第37卷，第1期，e2176618，2023年，doi：10.1080/08839514.2023.2176618。[17] D. Roselli、J. Matthews和N. Talagala，“管理人工智能偏见”，载于万维网会议（WWW）伴随论文集，美国旧金山，2019年，第539–544页，doi：10.1145/3308560.3317590。[18] L. H. Nazer等人，“人工智能算法中的偏见及其缓解建议”，PLOS Digit. Health，第2卷，第6期，e0000278，2023年6月，doi：10.1371/journal.pdig.0000278。[19] K. Chen、A. Shao、J. Burapacheep和Y. Li，“GPT-3如何回应不同公众对气候变化和Black Lives Matter的看法：对话式人工智能公平性的批判性评估”，arXiv:2209.13563 [cs.AI]，2022年9月。[在线]。网址：https://arxiv.org/abs/2209.13563[20] J. W. Gichoya等人，“人工智能陷阱及应避免的做法：缓解人工智能偏见”，英国放射学杂志，第96卷，第1150期，文章编号20230023，2023年10月，doi：10.1259/bjr.20230023。[21] Ç. Güven等人，“助力多样性与包容的人工智能”，arXiv:2501.05610 [cs.AI]，2025年1月。[在线]。网址：https://arxiv.org/abs/2501.05610[22] K. Kieslich、B. Keller和C. Starke，“设计即伦理的人工智能：评估公众对人工智能伦理设计原则重要性的认知”，arXiv:2106.00326 [cs.CY]，2021年6月。[在线]。网址：https://arxiv.org/abs/2106.00326[23] M. Whittaker等人，《残疾、偏见与人工智能》。美国纽约州纽约：AI Now Institute、纽约大学，2019年11月。[在线]。网址：https://ainowinstitute.org/publication/disability-bias-ai-2019[24] J. Seth，“人工智能的公众认知：情感倾向与机遇”，arXiv:2406.12640 [cs.CY]，2024年6月。[在线]。网址：https://arxiv.org/abs/2406.12640[25] A. Min，“人工智能与偏见：挑战、影响和补救措施”，国际科学研究杂志，第2卷，第11期，第3808–3817页，2023年10月。[26] S. G. Mayson，“偏见输入，偏见输出”，Yale Law J.，第128卷，第8期，第2122–2173页，2019年6月。[在线]。网址：https://yalelawjournal.org/article/bias-in-bias-out[27] R. A. Shams、D. Zowghi和M. Bano，“人人可用的人工智能：识别与多样性和包容相关的人工智能事件”，arXiv:2401.09045 [cs.CY]，2024年1月。[在线]。网址：https://arxiv.org/abs/2401.09045[28] W. McKinney，“Python中的统计计算数据结构”，载于第9届Python科学会议论文集，美国得克萨斯州奥斯汀，2010年，第56–61页。[29] C. R. Harris等人，“使用NumPy进行数组编程”，《自然》，第585卷，第7825期，第357–362页，2020年9月，doi：10.1038/s41586-020-2649-2。[30] P. Virtanen等人，“SciPy 1.0：Python科学计算的基础算法”，《自然·方法》，第17卷，第3期，第261–272页，2020年3月，doi：10.1038/s41592-019-0686-2。[31] E. Gazoni和C. Clark，“openpyxl——用于读取/写入Excel 2010 xlsx/xlsm文件的Python库”，2023年。[在线]。网址：https://openpyxl.readthedocs.io[32] Python软件基金会，“re——正则表达式操作”，Python 3文档，2023年。[在线]。网址：https://docs.python.org/3/library/re.html