# 机器学习中公平性属性的全球化：非洲卫生案例研究

**Mercy Nyamewaa Asiedu*** MASIEDU@GOOGLE.COM
*Google Research*

**Awa Dieng*** AWADIENG@GOOGLE.COM
*Google Research*

**Abigail Oppong** ABIGOPPONG@GMAIL.COM
*Ghana NLP*

**Maria Nagawa** MARIA.NAGAWA@DUKE.EDU
*Duke University*

**Sanmi Koyejo** SANMIK@GOOGLE.COM
*Stanford University, Google Research*

**Katherine Heller** KHELLER@GOOGLE.COM
*Google Research*

arXiv:2304.02190v1 [cs.LG] 5 Apr 2023

## 摘要

随着机器学习（ML）在医疗保健领域的应用日益广泛，人们开始呼吁关注机器学习的公平性，以理解并缓解这些系统可能带来的伦理问题。公平性对非洲的全球卫生具有重要影响，而非洲与全球南方之间本已存在不平等的权力失衡。本文旨在探讨全球卫生领域的公平性，并以非洲为案例研究。我们提出了在非洲语境中值得考虑的公平性属性，并阐述了这些属性在不同机器学习赋能的医疗模式中可能发挥作用的环节。本研究为进一步推进全球卫生公平性研究奠定基础，并呼吁采取行动。

## 1. 引言

机器学习（ML）模型有望对健康领域产生深远影响。然而，它们也可能传播反映现实世界中历史与当前不平等的偏差，并导致意料之外的有害结果，尤其是对弱势群体而言（Huang 等人，2022年；Haytham 和 Wang，2022年；Chen 等人，2021年；Char 等人，2018年；Gianfrancesco 等人，2018年；Obermeyer 等人，2019年）。偏差来源包括缺失数据、非代表性样本量、群体分布不均衡，以及分类错误或测量误差（Gianfrancesco 等人，2018年）。在健康领域，这可能会影响弱势群体——包括医疗护理不连续的患者、读写能力水平较低者、代表性不足的亚群体，以及/或使用测量设备或数据输入方式有限的临床设施的患者（Gianfrancesco 等人，2018年；Obermeyer 等人，2019年）。一个突出的例子是美国使用的一种算法，它依据有缺陷的成本变量，将病情更严重的黑人患者排在较低的护理优先级，从而延续了结构性社会经济差异（Obermeyer 等人，2019年）。这促成了算法公平性领域的兴起，并推动人们通过评估模型、了解可能导致机器学习模型作出不公平或有偏差决策的属性，尝试纠正机器学习偏差。

全球卫生有多种定义，但我们采用Beaglehole（2010年）提出的定义：“为促进全民健康而开展的跨国协作研究与行动”。公平性对全球卫生尤为重要。尽管取得了减少……等积极成果，全球卫生领域仍长期受到高收入国家（HICs）与低收入和中等收入国家（LMICs）之间权力失衡的不公平影响（Holst，2020年）。近期的全球卫生去殖民化运动揭示了“高收入国家生成的知识如何界定实践并影响思想，从而损害

* 这些作者贡献相同

© M.N. Asiedu、A. Dieng、A. Oppong、M. Nagawa、S. Koyejo 和 K. Heller。

低收入和中等收入国家的知识体系”（Eichbaum 等人，2021年）。Eichbaum 等人 （2021年）探讨了殖民主义、医学与全球卫生之间的交集，以及殖民主义如何持续影响全球卫生项目和合作伙伴关系。他们指出，即使使用“全球卫生”一词来描述低收入和中等收入国家的健康状况，也存在问题。鉴于大多数机器学习模型是在高收入国家利用其问题构建、人员、资源和数据开发的，并且可能在缺乏监管的情况下被引入低收入和中等收入国家，因此存在算法殖民主义的风险（Birhane，2020年；Mohamed 等人，2020年）这将进一步加剧当前全球卫生领域的权力失衡。另一方面，互联网接入的扩大，以及机器学习知识和工具的普及，为通过赋能研究合作伙伴关系开发适用于当地健康需求的机器学习工具，在全球卫生领域创造可持续变革带来了机遇。在探讨机器学习在健康领域可能如何有益于全球卫生时，我们也必须主动识别并减轻其可能造成的偏差。

迄今为止，卫生研究中的公平性推动了人们对机器学习开发流程中各环节偏差的理解与缓解（Chen 等人，2021年；Char 等人，2018年；Haytham 和 Wang，2022年）。然而，大多数研究都以高收入国家为背景，只有少数研究探讨了如何将公平机器学习的背景化应用拓展到西方以外。

Sambasivan 等人(2021年)采用混合方法研究，识别印度情境下公平机器学习所涉及的亚群体、代理指标和危害。对于自然语言处理应用，Bhatt 等人(2022年)识别出印度特有的差异维度，如地区和种姓，并在印度语境下重新界定全球性维度，如性别、宗教、性身份与性取向以及能力。Fletcher 等人(2021年)针对全球卫生中的公平性、偏差和人工智能的恰当使用提出了详细建议，并以印度的肺部疾病分类为例。

本文在先前研究的基础上，探讨全球卫生领域的公平性属性，并以非洲为案例：

- 我们识别出在涉及非洲时应予以考虑的公平性属性，并提出：（1）非洲与西方之间的差异维度，以及（2）非洲背景下的全球差异维度。
- 我们界定了健康领域机器学习的现实应用场景，在这些场景中应考虑这些维度。
- 最后，我们讨论了非洲背景下有效应用公平机器学习所面临的情境挑战以及尚待探索的研究方向。

尽管整个大陆可能存在一些共通主题，但非洲极为多元，其中的某些属性和差异维度可能适用于一些国家，却不适用于另一些国家。这项早期研究旨在奠定基础，以健康领域的机器学习应用为案例，探讨如何超越西方语境来理解公平性。我们希望借此传达这一研究领域的重要性，并激发人们对深入探索这一领域的兴趣。

## 2. 公平性考量

我们着重讨论全球尺度、大陆尺度和更地方尺度上的差异维度。我们有意不聚焦于非洲大陆上某个特定的“同质”地区，以凸显可能影响整个大陆的不公平，以及次区域和国家层面不同程度的不公平。这些差异维度是以非洲背景为依据界定的，其背后的逻辑也可应用于不同的地理区域。本节其余部分将讨论非洲与非非洲国家之间公平性的差异维度，以及非洲背景下公平性属性的考量。个体可能同时具有多个差异维度。关键术语定义见附录 5。

### 2.1. 非洲与西方国家之间的差异维度

**殖民历史：** 除埃塞俄比亚和利比里亚外，非洲所有国家都曾被欧洲人殖民。有充分证据表明，非洲的殖民历史及其造成的结构性挑战，至今仍在被殖民者与殖民统治者及受益者之间造成权力失衡（Eichbaum 等人，2021年）。多位学者提出，殖民历史是健康的社会决定因素之一（Turshen，1977年；Ramos 等人，2022年；Czyzewski，2011年）。这会对数据收集产生影响。我们提出，应将殖民历史视为一种公平性属性，因为它

可能导致历史偏差、代表性偏差、测量偏差、学习偏差、评估偏差和部署偏差，因此应在整个机器学习流程中予以考虑：问题构建、数据收集、模型开发和部署。尤其值得关注的是学习偏差，因为模型性能的客观指标不应局限于准确率，还应致力于改善影响指标。

**国家收入水平（NIL）：**殖民主义及其导致的结构性问题与国家收入水平相关。按国家收入水平划分，非洲国家数量为：低收入国家24个、中低收入国家17个、中高收入国家6个，以及高收入国家1个（Nada 等，2022年）。由于非洲大陆上的低收入和中等收入国家数量比例过高，由此产生的社会经济差异意味着，各国政府对健康领域机器学习研究的资金投入有限，也缺乏用于机器学习开发的临床数据生成设备和资源（Nabyonga-Orem 等，2021年）。

这导致问题定义更多地迎合外部资助要求，而非当地需求。这也会影响对数据生成工具的获取，进而造成代表性偏差，并导致机器学习开发所需的技术资源和人员资源有限。

**原籍国：**非洲分为五个不同的次区域：北非、西非、中非、东非和南非。这些地区的国家在国家收入水平、殖民历史、语言、文化、族裔以及种族亚群体方面各不相同。国界线是殖民产物，是瓜分非洲的争夺行动所造成的分割，埃塞俄比亚和利比里亚除外。这对族裔、资源和发展均有影响（Michalopoulos 和 Papaioannou，2016年）。不同国家可能采用不同的方法实施卫生策略。因此，在一个国家有效的机器学习应用未必能够迁移到另一个国家。应通过理解公平性影响可能因国家而异这一点，避免因采用一刀切模型而产生的聚合偏差。

### 2.2. 非洲背景下全球适用的差异维度

从外部引进机器学习模型时，应考虑这些差异，也应在大陆内部开发机器学习模型时加以考虑。大陆内部的差异可能导致不同类型的偏差。

**种族：**在公平性领域，种族通常指结构性种族主义。鉴于美国等国家的奴隶制历史及其后续种族差异，种族被视为一种社会建构，但这一概念并不直接适用于非洲。大多数非洲人（80%）是黑人，并受到全球反黑人言论的影响（Pierre，2013年）。从外部看来，北非阿拉伯人属于白人，南非的南非白人亦如此。非洲也有亚裔人群，主要分布在东非和南部非洲。

然而，大多数非洲人并不以种族认同自己，而更可能以国籍和族裔作为认同（Asante，2012年；Maghbouleh 等人，2022年）。人口健康问卷通常不会询问“种族”，但非洲裔人士可能更容易罹患某些疾病。如果在美国开发的模型在不同种族群体中表现良好，不能据此假定该模型在美国以外地区也能跨种族表现良好，因为差异背后的成因可能不同，并可能导致聚合偏差。

人们可以考察在全球尺度上影响非洲人的种族差异，也可以考察非洲不同国家内部的差异，例如北非黑人和白人（Pierre，2013年），或南非黑人和白人之间的差异（Verwey 和 Quayle，2012年）。这会影响健康服务可及性方面的社会经济差异、代表性偏差，以及基于种族的机器学习模型性能差异。

**族裔：**非洲拥有 3000 多个族群，是遗传多样性最丰富的大陆。与高收入国家不同，大多数非洲人认同某个族群。除了身体特征、语言和文化之外，族裔还定义了历史、抱负和世界观（Deng，1997）。

社会经济差异和获取医疗服务的方式可能沿着族裔界限呈现（Franck 和 Rainer，2012年）。这可能导致机器学习应用偏向优势族群，从而造成族裔不平等。

**宗教：**殖民主义之前，非洲人信奉传统宗教。随着西方宗教传入，基督教和伊斯兰教逐渐成为这片大陆上的主流宗教。有些

国家的基督教徒占多数，穆斯林和传统宗教信仰者占少数，反之亦然。宗教可能与族裔和社会经济地位相关，从而导致健康方面的“族裔宗教差异”（Gyimah 等人，2006年；Ha 等人，2014年）。宗教也可能影响人们获取、看待和采用健康实践的方式（White，2015年；Schmid 等人，2008年；Obasohan，2014年）。

**语言：**非洲有超过2000种语言，大多数非洲人会说不止一种（Heine，2000年）。非洲的口语通常以一种占主导地位的殖民语言或族群语言为主，例如英语、法语、阿拉伯语或葡萄牙语。通常有一种占主导地位的本土语言，以及若干影响力较小的地方方言。这些主导语言，尤其是殖民语言，有助于促进跨族裔交流，通常兼具书面语和口语形式，并用于教育和职业生活（Obondo，2007年）。使用较少见方言的人，以及受教育程度和读写能力水平较低的人，面临口语和书面语方面的语言障碍，这可能影响医疗服务提供中的沟通（Levin，2006年）。

**肤色：**在全球尺度上，非洲人的肤色无论深浅，都与黑人身份相关联，并会引发反黑人歧视（Pierre，2013年）。

在非洲黑人群体中，较浅的肤色可能被视为有助于获得更高的社会地位和更多优待（Nyoni-Kachambwa 等，2021年）。因此，大陆上有40%的人会进行美容性皮肤漂白（Asumah 等，2022年）。

漂白本身就是一项公共卫生问题，会增加出现严重健康问题的风险（Lewis 等人，2012年）。在健康领域机器学习中，这会影响从业者态度、皮肤病学（例如皮肤病检测），以及那些在深色皮肤上表现不佳的设备所造成的测量偏差。

**性别：** 历史上，受殖民遗产强化的非洲父权文化导致不同性别群体之间的财富分配不平等（Jaiyeola，2020年）。女性承担的家庭 + 育儿职责，以及近年增加的职业责任，进一步加重了负担，并降低了非洲女性获得基本医疗设施的机会（Judith 等人，2020年）。非洲的跨性别者和性别表达不符合传统规范者遭受污名和歧视，且很少有医疗项目为跨性别者提供服务（Luvuno 等人，2017年。这也会影响医疗服务可及性和治疗，尤其是针对高发性传播疾病的治疗（van der Merwe 等人，2022年）。

**性取向：**32个国家将同性恋定为犯罪，且在3个国家可判处死刑（Lars，2022年）。同性恋者受到污名化和歧视，导致他们在卫生系统中遭到排斥和边缘化（Sekoni 等人，2022年；Ross 等人，2021年）。

**读写能力与教育水平：**读写能力和教育水平会影响医疗护理可及性、求医行为以及对健康信息的理解，尤其是在信息以数字形式提供时（Amoah 和 Phillips，2018年）。这不仅关系到个人自身的健康，也关系到其子女的健康（Mwoya 和 Mpeta，2021年）。

**年龄：**年龄是全球公平性属性，在健康领域的机器学习模型开发和评估中均会用到，这是因为年龄特异性发病率和共病的存在。必须证明机器学习模型在不同年龄组中的性能公平（Mhasawade 等人，2021年）。

**城乡差距：**生活在农村地区的人群，其社会经济地位和读写能力往往较低，受教育程度也较低，且获取医疗机构服务的机会有限。这使农村地区的人群最容易受到不道德的机器学习实践的影响。然而，我们必须确保机器学习应用公平且有益，能够惠及农村地区，并改善健康差距。

**社会经济地位：**个人层面的社会经济差异是上述大多数差异维度的根源。非洲的财富分配差距位居全球第二（Seery 等人，2019年）。这可能导致健康领域的机器学习模型在部署可及性方面延续不公平，或主要对较贫困人群产生不准确的结果。

**残疾：**非洲人口中有 10% 至 20% 受到残疾影响。由于污名，以及国家层面资源和政策的不足，残疾可能会加剧下文列出的多数属性（Adugna 等人，2020年；McKinney 等人，2021年）。

**健康特定属性：**遗传和表型表现、已有的基础疾病，以及疾病表现的后期阶段，都可能影响模型性能。例如，在乳腺癌病例中，具有非洲血统的人群患三阴性乳腺癌的比例更高。在宫颈癌影像中，如果患者同时患有艾滋病病毒感染或宫颈炎等疾病，宫颈影像表现可能有所不同。这会直接影响不同健康特定属性下的表征偏差。

## 3. 按健康数据模态划分的情境障碍

健康领域机器学习依赖大型开源数据集。然而，这些数据集大多来自非洲以外的国家。虽然也有一些非洲健康数据集，但由于电子健康记录使用有限，这些数据集大多局限于调查。我们讨论不同健康数据模态中的公平性属性可能如何影响这些数据。偏差类型的定义见附录5。

**医学影像：** 机器学习在皮肤病学、病理学、乳腺摄影、超声、计算机断层扫描、磁共振成像和X射线等医学影像任务中的应用前景最为可观（Castiglioni等人，2021年）。然而，由于影像设备成本高昂且需要维护，非洲的医院普遍缺少这类设备。即使有影像设备，其质量和判读结果也可能参差不齐，导致来自高收入国家的机器学习模型泛化能力较差。如果有设备，也可能价格更高，且仅分布在城市地区，从而造成社会经济地位差异和城乡差距。此外，医学图像可能不会以数字形式存储，也可能未与其他患者健康记录关联。这会限制用于重新训练模型的数据，并可能造成表征偏差和评估偏差。

**调查数据：**人口与健康调查数据是非洲大陆规模最大且易于获取的纵向健康数据来源之一。这些调查用于提供不同地区和人口群体随时间变化的疾病趋势数据。尽管这些数据能够呈现不同背景的非洲人群，但在将其用于机器学习模型开发时，必须确保模型不会依赖代理指标或敏感属性，从而导致某些人口群体受到不公平对待。这会影响多个属性上的公平性，包括性别、性取向、族裔、残疾和社会经济地位。

**非结构化书面健康记录：** 针对非结构化健康记录的机器学习会运用大型计算机视觉和自然语言处理模型，提取症状和待办事项等信息，并提供疾病分类。这些模型通常使用大量资源过剩语言的文本数据进行预训练，因此已经表现出与文化、性别和种族相关的偏差。未经适当评估和微调就使用这些模型，可能会使已发现的偏差与语言相关偏差一同延续。除了建模方面的挑战，由于电子健康记录的可用性有限，医疗机构获取书面健康记录也受到限制。

**医疗语音：** 医疗专业人员会在各类医疗机构使用自动语音识别系统口述记录，无需因此占用患者护理时间。口音、说话方式（例如皮钦语）和读写能力可能会影响为非洲场景设计的语音识别算法。这会加剧语言差异，并间接影响族裔、原籍国、读写能力和教育。

**光学传感器设备：** 基于光学传感器的设备（如脉搏血氧仪和健身追踪器）在深色皮肤人群中的表现较差，这一点已有研究证实（C etal.，2022年）。为这些设备开发的机器学习模型可能会延续测量偏差。这会直接造成肤色差异，并可能间接关联到原籍国和族裔。

**面部识别：**面部识别算法已被用于健康领域，以诊断自闭症等疾病（Liu et al.，2016年）。然而，这些算法也被证明在肤色和性别方面存在较大的性能偏差（Buolamwini andGebru，2018年）。此外，不同文化间的面部表情也存在差异，必须予以考虑（Dailey etal.，2010年）。

**组学：** 利用机器学习从多组学数据中发现生物标志物，是一个快速发展的领域，具有潜力

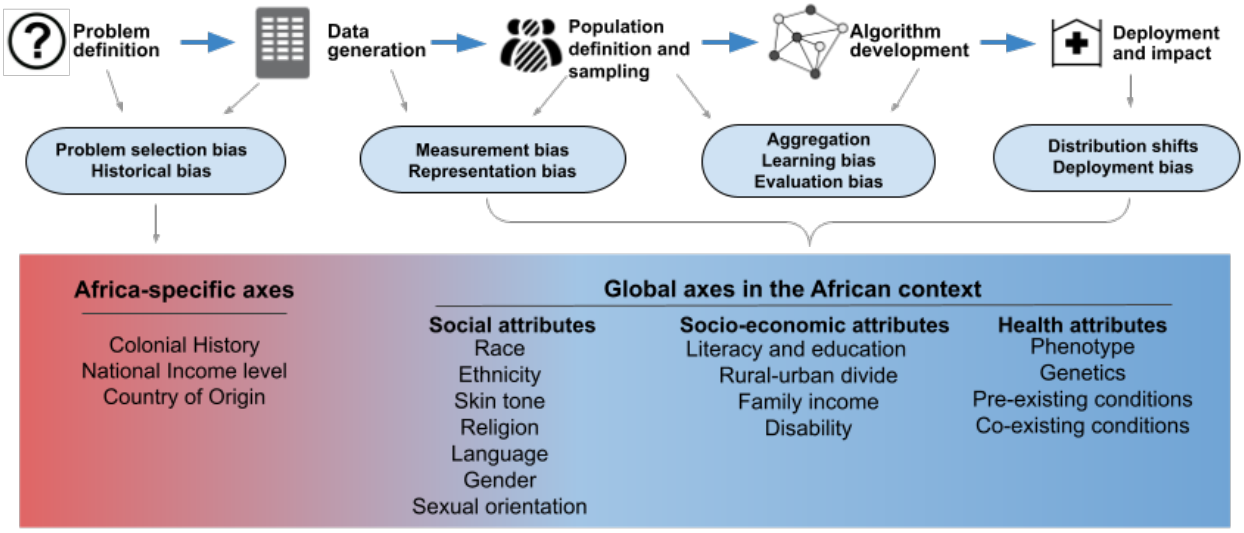



图 1：机器学习各环节中的偏差及其与非洲情境下差异维度的关联。

可用于精准诊断、预后和预测（S. 等，2021年）。非洲拥有丰富的遗传多样性，但现有组学数据代表性不足，仅占组学数据库的1%（Hamdi 等，2021年）。尽管在人类遗传与健康计划（H3Africa）联盟和H3Africa生物信息学网络（H3ABioNet）等联盟的推动下，数据量不断增加，但标注数据、族群多样性代表性，以及与患者健康史和个人资料的整合方面仍存在局限（Hamdi 等，2021年）。

**实验室检查指标：** 根据医院设施资源情况，患者可以在院内完成转诊安排的实验室检查，但通常会被转诊至外部实验室进行检测。经济条件较差的人群以及居住在农村地区、可能需要前往城市实验室机构的人群，获得实验室检测服务的机会尤其有限（Petti 等，2006年）。检测结果会交给患者，再由患者提供给转诊临床医生。由于电子健康记录有限，这些结果并不总是纳入患者的综合健康档案，导致健康信息彼此割裂。除数据获取受限外，用于解读实验室结果的基础设施和人员（例如病理学家）也十分有限影响诊断效率和准确率（Petti 等，2006年）。这些因素给机器学习带来了独特挑战，而这类挑战在高收入国家并不存在。

这些基于模态的情境障碍会影响分类、医学图像分割、自动语音识别、自然语言处理和健康推荐系统等上游机器学习任务，而这些任务可能缺乏针对非洲的数据集和情境信息。

## 4. 对机器学习的启示与研究机会

为解决机器学习模型中的偏差，研究人员提出了多种缓解策略。迄今为止，大多数方法学进展都聚焦于：（1）定义模型应满足的公平性准则；（2）在模型开发流程的各个环节设计相应程序，以确保满足该准则。鉴于这些步骤与具体问题本身存在内在关联，了解在非洲背景下构建和部署模型可能面临哪些具体挑战至关重要（Behara等人，2022年；Waljee等人，2022年）。

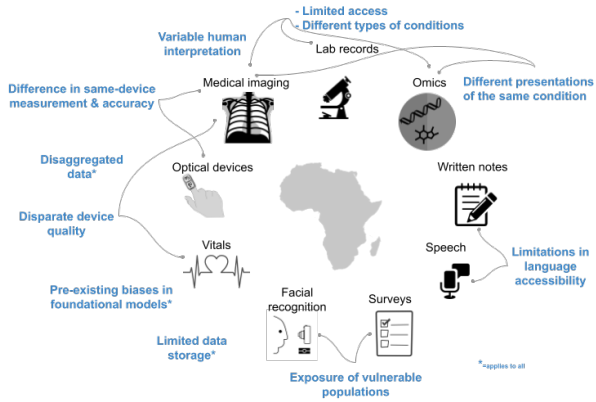



图2：按卫生服务类别划分的健康领域机器学习所面临的非洲语境化障碍。*=适用于所有类别

**公平性标准的情境化：**人口统计均等和机会均等等常见算法公平性指标（Barocas等人，2019年）最初的形式化定义，与美国（Barocas和Selbst，2016年）针对差别影响和差别待遇的反歧视法有着内在联系。这推动了相关方法学的发展，以确保自动化系统不会因种族、性别等与具体情境相关的受保护属性而产生歧视性影响（Hutchinson和Mitchell，2018年）。正如前文所述，高收入国家与低收入和中等收入国家的多样性，以及受保护属性之间的差异维度，使得研究哪些公平性定义适用于当地历史背景成为一项根本需求。由于这一领域在非洲大陆仍处于起步阶段，因此存在一个独特的机遇，可让不同利益相关方（研究人员、政策制定者、治理机构等）共同根据当地法律与信念界定相关的公平性目标。在这一阶段采用跨学科方法尤为重要，可避免因公平性定义相互冲突（Verma和Rubin，2018年；Kleinberg等人，2016年；Chouldechova，2017年）以及纯数学形式化方法而导致所有群体的性能意外下降（Mittelstadt等人，2023年）。

**公平性的实践考量：**算法公平性领域的后续研究已重点指出，偏差可能在机器学习开发流程的每个阶段以不同方式显现，尤其是——在以下阶段：(1) 问题选择，(2) 数据收集与处理，(3) 算法开发与评估，以及 (4) 部署后考量（Chen 等人，2021年）。文献中提出了大量去偏方法，以应对第 (3) 至 (5) 阶段可能出现的偏差（Mehrabi 等人，2021年；Barocas 等人，2019年）。除了更深入地了解如何调整这些缓解策略，以兼顾低收入和中等收入国家特有的不同公平性属性外，在非洲卫生领域应用机器学习技术的一大障碍，还在于问题表述、数据收集和数据利用（Okolo 等人，2023年）。下文将举例说明在非洲使用机器学习工具所面临的这些挑战，以及需要重点考虑的公平性问题。

- 问题选择：正如（Chen 等人，2021年）所指出的，外部组织之间不一致的激励所导致的偏差可能会严重影响例如哪些疾病得到研究。这会连带影响所收集的数据，而不论这些数据是否与当地人口相关。在这一步，尤其需要考虑利益相关方可能存在的偏差，并确保受影响社区参与决策。

- 数字化健康记录匮乏：由于缺乏基础设施和计算资源来保存数字副本，患者的健康记录仍主要以纸质形式保存。农村地区这一现象尤为严重，导致不平等进一步加剧。认识到电子记录对于利用机器学习工具至关重要，促使人们提出新的方案，将现有的手写数据数字化。这一步面临的公平性问题包括：如何开发能够可靠处理服务不足地区非洲语言的自动化工具，以及如何确保此类工作在各地区得到均衡开展。

- 调查数据的普及：机器学习在健康领域的大多数应用仍依赖于在临床环境中收集的数据（Mhasawade 等人，2021年）。在非洲背景下，自我报告调查数据的普及带来了多种偏差，包括因受访者而导致的群体代表性失衡，以及

关于受访者属性的自我报告信息可能不准确。将机器学习方法应用于此类数据模态，以及如何处理测量偏差，仍是一个研究不足的领域。

**使用预训练模型时需谨慎：**鉴于来自非洲、可随时获取的大型训练数据集十分稀缺，一种常见做法是对大型预训练模型进行微调，以适应下游任务。这种做法尤其诱人，因为它能让资源和基础设施较为受限的环境借助大型模型的强大能力，而训练这些模型需要大量资源和基础设施。然而，这种做法存在若干隐患，可能导致最终模型出现意料之外的偏差。

一个普遍的担忧是，模型在部署现成模型或针对目标数据集进行微调时，可能会遇到分布偏移，而公平性属性无法适应这种变化。然而，即使在训练时采用了偏差缓解技术，也无法保证公平性属性在部署后或微调后仍能保持。下文将举例说明非洲背景下现实世界中的一些分布偏移。

- 人口统计偏移是指公平性属性的分布在训练和测试阶段之间发生变化。将基于西方中心数据训练的模型部署到非洲数据时，很可能出现这种情况；例如，两个数据集中的种族群体构成可能不同。

- 协变量偏移是指训练数据与测试数据之间的输入特征分布发生变化。例如，基于机器学习的皮肤病学诊断模型可能是在较浅肤色的数据上训练的，因此可能无法识别深色皮肤上同一种疾病的表现。另一个例子是疾病出现的阶段。由于预防性筛查有限，疾病确诊时可能比在高收入国家更晚，从而影响模型分类的数据分布。

- 标签偏移是指训练数据与测试数据之间结果分布的变化。这类偏移可见于健康状况的差异，例如某些疾病可能仅见于非洲，且在西方国家的训练数据集中采样不足。

确保公平性属性在不同领域之间以及分布偏移情况下的泛化能力和可迁移性，是一个新兴且活跃的研究领域（Schrouff 等人，2022年；Giguere 等人，2022年；Singh 等人，2021年；Baldini 等人，2022年；Sadeghi 和 Boddeti，2020年）。非洲背景下的具体应用案例可以作为启发性示例，推动这一领域取得具有实际影响的进展。

虽然本节讨论的是基于机器学习的缓解策略，但解决机器学习偏差需要从非洲医疗保健领域的结构性和系统性变革着手。确保人人都能获得基本医疗保健；增加政府对医疗服务可及性、以健康为基础的创业和健康相关研究的投入，以及私人投资；以可持续且符合当地背景的方式开发和部署健康工具和技术，这些都是缩小健康差距的关键。

## 5. 结论

机器学习为非洲带来积极影响、推动全球健康公平提供了独特机遇。然而，我们应主动采取措施，防止危害、减少偏差并确保公平性。要开发公平的机器学习模型，首先需要了解特定情境下的公平性属性，以及应在何处应用这些属性。我们希望这项工作能够成为起点，为全球卫生领域的机器学习公平性奠定基础，并聚焦非洲。未来的工作将对机器学习模型进行定性和定量分析，考察它们对这些属性的响应，并提出缓解方案。

## 参考文献

Molalign B. Adugna、Fatima Nabbouh、Selvia Shehata 和 Setareh Ghahari。《撒哈拉以南非洲低收入和中等收入国家残障儿童获得医疗保健的障碍与促进因素：一项范围综述》。*BMC*《卫生服务研究》，20:1–11，2020年。

Padmore Adusei Amoah 和 David R. Phillips。健康素养与健康：重新思考加纳全民健康覆盖的策略。《公共卫生》，159：40–49，2018年。

Godfried Agyeman Asante。在美国成为“黑人”：探索非洲移民的种族身份发展。《明尼苏达州立大学曼凯托分校》，2012年。

Mubarick N. Asumah、Abubakari Abdulai、Edem K. Dzantor、Valentine Ayamgba、Ayishetu Gariba、Grace Buremah-Excellence 和 Doreen-Remember Donkor。加纳年轻成年人皮肤漂白的患病率及相关因素。公共卫生毒理学，2.1:1–9，2022年。

Ioana Baldini、Dennis Wei、Karthikeyan Nate- san Ramamurthy、Moninder Singh 和 Mikhail Yurochkin。公平性可能因你而异：预训练语言模型在有毒文本分类中的公平性。载于计算语言学协会论文集：*ACL 2022*，第 2245–2262 页，爱尔兰都柏林，2022年5月。计算语言学协会。数字对象标识符：10.18653/v1/2022. findings-acl.176。网址 https://aclanthology.org/2022.findings-acl.176。Solon Barocas 和 Andrew D. Selbst。大数据的差别影响。《加利福尼亚法律评论》，104 (3):671–732，2016年。ISSN 00081221。网址http://www.jstor.org/stable/24758720。Solon Barocas、Moritz Hardt 和 Arvind Narayanan。《公平性与机器学习：局限与机遇》。fairmlbook.org，2019年。http://www.fairmlbook.org。Ruth Bonita Beaglehole、Robert。什么是全球卫生？*Global Health Action*，3，2010年。Kavita Behara、Ernest Bhero、John Terhile Agee 和 Vinay Gonela。医学诊断中的人工智能：南非背景下的综述。*Scientific African*，17:e01360，2022年。ISSN 2468-2276。数字对象标识符：https://doi.org/10.1016/j.sciaf. 2022.e01360。网址 https://www.sciencedirect.com/science/article/pii/S2468227622002654。Shaily Bhatt、Sunipa Dev、Partha Talukdar、Shachi Dave 和 Vinodkumar Prabhakaran。在自然语言处理领域重新审视公平性：印度案例。*arXiv* 预印本，2022年。数字对象标识符：arXiv:2209.12226。Abeba Birhane。非洲的算法殖民化。*SCRIPTed*，17.2:389，2020年。数字对象标识符：10.2966/scrip. 170220.389。

Joy Buolamwini 和 Timnit Gebru。《Gender Shades：商业性别分类中的交叉性准确率差异》，发表于公平、问责与透明度会议。*PMLR*，第 77–91 页，2018年。Shi C、Goodall M、Dumville J、Hill J、Norman G、Hamer O、Clegg A、Watkins CL、Georgiou G、Hodkinson A 和 Lightbody CE。《按皮肤色素沉着程度测量血氧饱和度时脉搏血氧测定的准确性：系统综述与荟萃分析》。《*BMC*医学》，20，2022年。Isabella Castiglioni、Leonardo Rundo、Marina Codari、Giovanni Di Leo、Christian Salvatore、Matteo Interlenghi、Francesca Gallivanone、Andrea Cozzi、Natascha Claudia D’Amico 和 Francesco Sardanelli。《人工智能在医学图像中的应用：从机器学习到深度学习》。《医学物理学》，83:9–24，2021年。Danton S. Char、Nigam H. Shah 和 David Magnus。《在医疗保健中实施机器学习——应对伦理挑战》。《新英格兰医学杂志》，378.11:981，2018年。Irene Y. Chen、Emma Pierson、Sherri Rose、Kadija Ferryman Shalmali Joshi 和 Marzyeh Ghassemi。《医疗保健领域的伦理机器学习》，《生物医学数据科学年度评论》。《新英格兰医学杂志》，4:123–144，2021年。Alexandra Chouldechova。《具有差异影响的公平预测：累犯预测工具中的偏差研究》，2017年。网址：https://arxiv.org/abs/1703.00056。Karina Czyzewski。《殖民主义：更广泛的健康社会决定因素》。《国际原住民政策期刊》，2.1，2011年。Matthew N. Dailey、Carrie Joyce、Michael J. Lyons、Miyuki Kamachi、Hanae Ishi、Jiro Gyoba 和 Garrison W. Cottrell。《面部表情识别中文化差异的证据及计算解释》。《情绪》，10.6:874，2010年。Francis M. Deng。《族裔：非洲困境》。《布鲁金斯评论》，15.3:28–31，1997年。Cynthia Dwork、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Richard Zemel。《通过认知实现公平》，载于第三届理论计算机科学创新会议论文集，ITCS ’12，第214–226页，纽约，

美国纽约，2012年。计算机协会。国际标准书号9781450311151。数字对象标识符：10.1145/2090236.2090255。网址：https://doi.org/10.1145/2090236.2090255。

Quentin G. Eichbaum、Lisa V. Adams、Jessica Evert、Ming-Jung Ho、Innocent A. Semali 和 Susan C. van Schalkwyk。去殖民化全球卫生教育：重新思考机构合作伙伴关系与方法。《学术医学》，96.3:329–335，2021年。

Richard Rib´on Fletcher、Audace Nakeshimana 和 OlusubomiOlubeko。探讨人工智能和机器学习在全球卫生领域中的公平性、偏差及恰当应用。《人工智能前沿》，3，2021年。

Raphael Franck 和 Ilia Rainer。领导者的族裔是否重要？撒哈拉以南非洲的族裔偏袒、教育与健康。《美国政治学评论》，106.2:294–325，2012年。

Milena A. Gianfrancesco、Suzanne Tamang、Jinoos Yazdany 和GabrielaSchmajuk。使用电子健康记录数据的机器学习算法中可能存在的偏差。《美国医学会杂志·内科学》，178.11:1544–1547，2018年。

Stephen Giguere、Blossom Metevier、Yuriy Brun、Philip S. Thomas、ScottNiekum 和 Bruno Cas- tro da Silva。《人口统计分布偏移下的公平性保障》。载于 国际学习表征会议，2022年。网址 https://openreview.net/forum?id=wbPObLm6ueA。

Stephen Obeng Gyimah、Baffour K. Takyi 和 Isaac Addai。《非洲女性生殖健康需求面临的挑战：宗教与加纳孕产妇卫生服务利用》。《*Social Science & Medicine*》，62.12：2930–2944，2006年。

Wei Ha、Peter Salama、Stanley Gwavuya 和 Chi- f undo Kanjala。《宗教是否是妇幼健康中被忽视的变量？来自津巴布韦的证据》。《*Social Science & Medicine*》，118：80–88，2014年。

Y. Hamdi、L. Zass、H. Othman、F. Radouani、I. Al- lali、M. Hanachi、C.J. Okeke、M. Chaouch、M.B. Tendwa、C. Samtal 和 R. Mohamed Sallam。非洲的人类组学与计算生物学研究：当前挑战与前景。《*OMICS*：整合生物学杂志》，25.4：213–233，2021年。

Moritz Hardt、Eric Price、Eric Price 和 Nati Srebro。监督学习中的机会均等。载于 D. Lee、M. Sugiyama、U. Luxburg、I. Guyon 和 R. Garnett 主编的《神经信息处理系统进展》，第 29 卷。Curran Associates, Inc.，2016年。网址 https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf。

Suresh Harini 和 John Guttag。《理解机器学习生命周期中潜在的危害来源》。*MIT Case Studies in Social andEthical Responsibilities of Computing*，8，2021年。

Siala Haytham 和 Yichuan Wang。《将人工智能转变为负责任的医疗保健工具：系统综述》。*Social Science & Medicine*，第114782页，2022年。

Bernd Heine。《非洲语言：导论》。剑桥大学出版社，2000年。

Jens Holst。全球卫生——兴起、霸权趋势与生物医学还原论。”全球化与卫生。*Global Health Action*，16.1:1–11，2020年。

Jonathan Huang、Galal Galal、Mozziyar Etemadi 和 Mahesh Vaidyanathan。临床机器学习模型中种族偏见的评估与缓解：范围综述。*JMIR Medical Informat-ics*，10.5:1544–1547，2022年。

Ben Hutchinson 和 Margaret Mitchell。测试（不）公平性的50年：机器学习的经验教训。*CoRR*，abs/1811.10104，2018年。网址http://arxiv.org/abs/1811.10104。

E. O. Jaiyeola. 父权制与殖民化：尼日利亚性别不平等的“育婴室”。*Journal of Research on Women and Gender*，10.1：3–22，2020年。

Anaman-Torgbor Judith、Seth Kwadjo Angmorterh、Dzifa Dordunoo 和 Eric Kwasi Ofori。宫颈癌筛查行为与挑战：撒哈拉以南非洲视角。*Pan African MedicalJournal*，36.1，2020年。

Jon M. Kleinberg、Sendhil Mullainathan 和 Manish Raghavan。风险评分公平判定中的内在权衡。*CoRR*，abs/1609.05807，2016年。网址 http://arxiv.org/abs/1609.05807。

Matt J Kusner、Joshua Loftus、Chris Russell 和 Ricardo Silva。反事实公平。在 I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编者的神经信息处理系统进展，第 30 卷。Curran Associates, Inc.，2017年。网址
h
t
t
ps://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf。

Kamer Lars。《非洲同性关系刑事化》，2020年、2022年。网址
https://www.statista.com/statistics/1269999/criminalization-of-same-sex-relations-in-africa/。

Michael E. Levin。语言作为获得医疗护理的障碍：讲科萨语的患者在南非一家儿科教学医院的就医情况。《南非医学杂志》，96.10:1076–1079，2006年。

Kelly M. Lewis、Karie Gaska、Navit Robkin、Amber Martin、Emily Andrews 和 Justin Williams。预防皮肤漂白干预措施的必要性：聚焦坦桑尼亚。《黑人研究杂志》，43.7: 787–805，2012年。

Wenbo Liu、Ming Li 和 Li Yi。基于面孔加工异常识别自闭症谱系障碍儿童：一种机器学习框架。《自闭症研究》，9.8:888–898，2016年。

Zamasomi Luvuno、Busisiwe Ncama 和 Gugu Mchunu。医疗卫生工作者对跨性别患者治疗与护理的知识、态度和实践：南非夸祖鲁-纳塔尔省的一项定性研究。《性别与行为》，15.2:8694–8706，2017年。

Neda Maghbouleh、Ariela Schachter 和 Ren´e D. Flores。中东和北非裔美国人可能不会被视为白人，也可能不认为自己是白人。《美国国家科学院院刊》，119.7:547–551，2022年。

Emma L. McKinney、Victor McKinney 和 Leslie Swartz。南非残障人士的医疗服务可及性：任何时候都不理想，COVID-19 疫情期间更糟？《南非家庭医学》，63.3:116–128，2021年。

NinarehMehrabi、FredMorstatter、NripsutaSaxena、Kristina Lerman 和 Aram Galstyan。关于机器学习中的偏见与公平性的综述。*ACM* 计算机调查，54(6)，2021年7月。ISSN 0360-0300。数字对象标识符：10.1145/3457607。网址https://doi.org/10.1145/3457607。

Vishwali Mhasawade、Yuan Zhao 和 Rumi Chu- nara。机器学习与公共健康和人群健康中的算法公平性。《自然·机器智能》，3.8：116–128，2021年。Stelios Michalopoulos 和 Elias Papaioannou。瓜分非洲的长期影响。美国经济评论，7：1802–1848，2016年。

Brent Mittelstadt、Sandra Wachter 和 Chris Russell。公平机器学习中的不公平：默认采取降低标准和严格平等主义。2023年。网址
https://ssrn.com/abstract=4331652。

Shakir Mohamed、Marie-Therese Png 和 William Isaac。《去殖民化人工智能：去殖民化理论作为人工智能中的社会技术前瞻》。*Phi-losophy & Technology*，33：659–684，2020年。

Byaro Mwoya 和 Daniel Mpeta。中等教育及其对儿童健康的影响：来自撒哈拉以南非洲的实证证据。《非洲经济评论》，9.2：116–128，2021年。

Juliet Nabyonga-Orem、James Avoka Asa mani 和 Micheal Makanga。非洲卫生研究治理现状：我们了解什么，以及如何改进？《卫生研究政策与系统》，19:1–14，2021年。

Hamedeh Nada、Catherine van Rompaey、Eric Metreau 和 Schweta Grace Eapen。《世界银行按收入水平划分的最新国家分类：2022—2023年》，2022年。

Princess Nyoni-Kachambwa、Wanapa Naravage、Nigel F. James 和 Marc Van der Putten。津巴布韦女性皮肤漂白及其相关因素的初步研究。《非洲健康科学》，11.1:132–9，2021年。

Phillips E Obasohan。尼日利亚育龄女性的宗教、族裔与避孕措施使用。《国际母婴健康与艾滋病杂志》，3.1:63–73，2014年。

Ziad Obermeyer、Brian Powers、Christine Vogeli 和 Sendhil Mullainathan。《剖析一种用于管理人群健康的算法中的种族偏见》。《科学》，366:6464，2019年。

M.A Obondo。《后殖民时期非洲英语教学与母语教学之间的张力》。载于《英语语言教学国际手册》。《施普林格国际教育手册系列》，2007年。

Chinasa T. Okolo、Kehinde Aruleba 和 George Obaido。《非洲负责任人工智能——挑战与机遇》，第35–64页。施普林格国际出版公司，查姆，2023年。国际标准书号 978-3-031-08215-3。数字对象标识符：10.1007/978-3-031-08215-3 3。网址 https://doi.org/10.1007/978-3-031-08215-3_3。

Cathy A. Petti、Christopher R. Polage、T homas C. Quinn、Allan R. Ronald 和 Merle A. Sande。《非洲实验室医学：有效医疗保健的一项障碍》。《临床传染病》，42.3:377–382，2006年。

Jemima Pierre。《当今非洲的种族问题：一篇评论》。《文化人类学》，28.3:547–551，2013年。Jos´e G.P´erez Ramos、Adriana Garriga-L´opez 和 Carlos E. Rodr´ıguez-D´ıaz。《殖民主义如何成为波多黎各健康的社会结构性决定因素？》《美国医学会伦理学杂志》，24.4:305–312，2022年。

Michael W. Ross、John Kashiha、Lourence Misedah、Lucy R. Mgopa、Markus Larsson、Anette Agardh 和 Krishna Kumar Venkitachalam。《回应非洲男男性行为者的医疗保健需求：撒哈拉以南非洲医疗保健障碍及促进艾滋病病毒和性传播感染治疗》。《东非健康与科学杂志》，3.1:59–77，2021年。Reel Parminder S.、Smarti Reel、Ewan Pearson、Emanuele Trucco 和 Emily Jefferson。《运用机器学习方法进行多组学数据分析：综述》。《生物技术进展》，49，2021年。Bashir Sadeghi 和 Vishnu Naresh Boddeti。《为预训练的偏见表征赋予公平性》。载于《*2020*年*IEEE/CVF*计算机视觉与模式识别会议研讨会（*CVPRW*）》，第75–82页，2020年。数字对象标识符：10.1109/ CVPRW50498.2020.00016。

Nithya Sambasivan、Erin Arnesen、Ben Hutchinson、Tulsee Doshi 和 Vinodkumar Prabhakaran。《重新构想印度及其他地区的算法公平性》。载于《*2021*年*ACM*公平性、问责制与透明度会议论文集》，第315–328页，2021年。

B. Schmid、E. Thomas、J. Olivier 和 J. R. Cochrane。宗教实体对撒哈拉以南非洲健康的贡献。*OpenUCT*，2008年。Jessica Schrouff、Natalie Harris、Oluwasanmi O Koyejo、Ibrahim Alabdulmohsin、Eva Schnider、Krista Opsahl-Ong、Alexander Brown、Subhra- jit Roy、Diana Mincu、Chrsitina Chen、Awa Di- eng、Yuan Liu、Vivek Natarajan、Alan Karthike- salingam、Katherine A Heller、Silvia Chiappa 和 Alexander D'Amour。诊断真实世界医疗场景中分布偏移下公平性迁移失效的问题。载于 Alice H. Oh、Alekh Agar- wal、Danielle Belgrave 和 Kyunghyun Cho 编，《神经信息处理系统进展》，2022年。网址 https://openreview.net/forum?id=K-A4tDJ6HHf。

E. Seery、J. Okanda 和 M Lawson。两大洲的故事：在非洲抗击不平等。乐施会，2019年。数字对象标识符：10.21201/201.4719。Adekemi Oluwayemisi Sekoni、Kate Jolly 和 Nicola Kay Gale。隐性医疗保健人群：运用交叉性理论化尼日利亚（非洲）LGBT+人群的经历。《全球公共卫生》，17.1:134–149，2022年。Harvineet Singh、Rina Singh、Vishwali Mhasawade 和 Rumi Chunara。协变量偏移下的公平性违规及缓解措施。载于*2021*年*ACM*公平性、问责制与透明度会议论文集，FAccT '21，第3–13页，纽约，纽约州，美国，2021年。计算机协会。国际标准书号 9781450383097。数字对象标识符：10. 1145/3442188.3445865。网址 https://doi.org/10.1145/3442188.3445865。

Meredeth Turshen。殖民主义对坦桑尼亚健康和卫生服务的影响。《国际卫生服务杂志》，7.1:7–35，1977年。van der Merwe、L. Leigh-Ann、Allanise Cloete 和 Donald Skinner。如何利用艾滋病毒倡导确保优质的跨性别者医疗保健：来自南非的经验。《柳叶刀》，399.10330:1099– 1101，2022年。Sahil Verma 和 Julia Rubin。公平性定义详解。载于国际软件公平性研讨会论文集，FairWare '18，第1–7页，纽约，纽约州，美国，2018年。计算机协会。国际标准书号9781450357463。数字对象标识符：10.

1145/3194770.3194776. 网址 https://doi.org/10.1145/3194770.3194776.

C. Verwey 和 M. Quayle。《白人性、种族主义与后种族隔离时代南非的南非白人身份》。《非洲事务》，111:551–575，2012年。

Akbar K Waljee、Eileen M Weinheimer-Haus、A m- ina Abubakar、Anthony K Ngugi、Geoffrey H Siwo、Gifty Kwakye、Amit G Singal、Arvind Rao、Sameer D Saini、Andrew J Read、Jessica A Baker、Ulysses Balis、Christopher K Opio、Ji Zhu 和 Mansoor N Saleh。人工智能与机器学习在撒哈拉以南非洲结直肠癌早期检测和诊断中的应用。《*Gut*》，71(7):1259–1265，2022年。ISSN 0017-5749。数字对象标识符：10.1136/gutjnl-2022-327211。网址 https://gut.bmj.com/content/71/7/1259。

Peter White。加纳非洲传统宗教中的疾病与医疗保健概念。《*HTS: The- ological Studies*》，71.3:1–7，2015年。

## 附录A. 定义

本文定义了文中使用的术语。虽然公平性尚无标准化定义，但我们主要遵循（Mehrabi 等人，2021年）中描述的分类体系。我们还提供其他术语，以帮助读者进一步理解公平性。

**歧视：**‘歧视可被视为不公平的一种来源，它源于基于敏感属性的人类偏见和刻板印象，且可能是有意或无意的’。

**偏差：**‘偏差可被视为不公平的一种来源，它源于数据收集、抽样和测量。’

**偏差类型** (Harini 和 Guttag，2021年)

历史偏差：‘即使数据测量和抽样都完全准确，只要现实世界（无论是现在还是过去）导致模型产生有害结果，历史偏差就会出现。这样的系统即使准确反映了现实世界，仍可能对某个人群造成伤害。评估历史偏差时，通常需要考量其对特定群体造成的表征伤害（例如强化刻板印象）’。

代表性偏差：‘当开发样本对人群中的某些部分代表不足，进而无法很好地推广到使用人群中的某个子集时，就会出现代表性偏差’。

测量偏差：‘当为预测问题选择、收集或计算特征和标签时，就会产生测量偏差。通常，特征或标签是一个代理指标（具体测量），用于近似表示某种未被直接编码或观测到的构念（想法或概念）。当代理指标过度简化、在不同群体中的测量方式不同，或准确率因群体而异时，代理指标就会产生问题’。

聚合偏差：‘当数据中存在应区别对待的潜在群体或样例类型，却对所有数据一概采用同一种模型时，就会产生聚合偏差。特定数据集可能代表着背景、文化或规范各异的个人或群体，同一个变量在不同群体中可能意味着截然不同的东西。聚合偏差可能导致模型对任何群体都不是最优的，或者模型只适用于占主导地位的人群’。

学习偏差：‘当建模选择扩大了数据中不同样例之间的性能差异时，就会产生学习偏差。若优先考虑某一目标（例如总体准确率）会损害另一目标（例如差别影响），就可能出现问题’。

评估偏差：‘当用于特定任务的基准数据无法代表实际使用情况时，就会产生评估偏差’。

部署偏差：‘当模型旨在解决的问题与其实际使用方式不一致时，就会产生部署偏差’。

**公平性**：‘在决策语境中，公平性是指不因个人或群体固有或后天形成的特征而对其有所偏见或偏袒’。

**公平性类型**

个体公平性 (Dwork 等人,2012年)：‘个体公平性体现为这样一项原则：对于特定任务而言，任何两个相似的个体都应被归为相似类别

特定任务应以类似方式进行分类'。

反事实公平 (Kusner 等人，2017年)「反事实公平的定义基于这样一种'直觉'：如果一个人属于不同人口群体时，在现实世界和反事实世界中得到的决策相同，那么该决策对这个人就是公平的。」

群体公平：
广义而言，群体公平理念旨在「平等对待不同群体」。以下介绍文献中两种主要的统计群体公平定义。

- 机会均等 (Hardt 等人，2016年)：「这一公平性理念要求，正类中的个体被判定为正向结果的概率，在受保护群体和非受保护群体（女性和男性）成员之间相等。换言之，机会均等定义指出，受保护群体和非受保护群体应具有相同的真正率。」

- 人口统计均等 (Dwork 等人，2012年)：「要求在受保护群体中被预测为正类（或负类）的个体总体比例，与全体人群中的相应比例相同。」

关系领域公平：'一种能够捕捉某一领域中关系结构的公平性概念——不仅考虑个体的属性，还考虑个体之间的社会、组织及其他联系。'

子群体公平： 子群体公平旨在兼顾群体公平与个体公平的优点。它选取一项群体公平约束，例如使假阳性率相等，并检验该约束是否适用于大量亚群体。

**差异维度**

非洲特有的差异维度： 主要影响非洲的差异。

非洲语境下的全球差异维度：具有全球影响、但结合非洲具体语境加以阐释的差异。