深度神经网络公平性缺陷的信息论测试与调试

Verya Monjezi
vmonjezi@miners.utep.edu德克萨斯大学埃尔帕索分校

阿舒托什·特里维迪
ashutosh.trivedi@colorado.edu科罗拉多大学博尔德分校

谭刚gtan@psu.edu宾夕法尼亚州立大学

Saeid Tizpaz-Niari
saeid@utep.edu德克萨斯大学埃尔帕索分校

arXiv:2304.04199v1 [cs.SE] 9 Apr 2023

摘要——**深度前馈神经网络（DNN）日益广泛地应用于对社会经济至关重要的决策支持软件系统中。DNN 尤其擅长从训练数据中发现最小且充分的统计模式。因此，DNN 可能学会编码决策——放大现有偏差或引入新的偏差——从而使受保护的个人或群体处于不利地位，并可能违反法律保护。尽管现有基于搜索的软件测试方法在发现公平性缺陷方面成效显著，但它们并未为这些缺陷提供调试辅助信息，例如严重程度和因果解释，而这些信息对于帮助开发人员确定问题优先级并决定后续行动至关重要。我们能否衡量 DNN 中公平性缺陷的严重程度？这些缺陷是否表明训练不当，还是仅仅反映了训练数据中存在的偏差？为回答这些问题，我们提出 DICE：一个用于发现和定位 DNN 中公平性缺陷的信息论测试与调试框架。**

DICE 的主要目标是通过按照公平性缺陷的严重程度对其排序，协助软件开发人员对公平性缺陷进行分类处理。为实现这一目标，我们根据决策中使用的受保护信息（以比特为单位）来量化公平性。对公平性缺陷进行定量分析不仅有助于对这些缺陷进行排序，我们的实证评估还表明，由此产生的搜索空间平滑性提高了搜索效率。在定量公平性的指导下，我们提出了一个因果调试框架，用于定位导致公平性缺陷的训练不足的层和神经元。我们在十个面向社会关键任务开发的深度神经网络上开展的实验表明，DICE 能够高效地刻画歧视程度，有效生成歧视性实例（相较于最先进技术），并定位具有显著偏差的层/神经元。

I. 引言

人工智能辅助的软件解决方案——越来越多地以深度神经网络 [1] (深度神经网络s)的形式实现——已经大举进入关键软件基础设施领域，在这些领域中，它们经常协助进行社会经济和法律关键性的决策 [2]。此类人工智能辅助软件的实例包括决定累犯风险的软件、预测福利资格的软件，以及决定是否审计某位纳税人的软件。基于深度神经网络的软件开发由信息瓶颈原理 [3], 驱动，涉及在过拟合与检测有用且简约的模式之间进行微妙的平衡。因此，此类解决方案编码并放大既有偏差也就不足为奇了训练数据。更糟糕的是，不当的训练甚至可能引入训练数据中不存在或与决策无关的偏差。由此产生的公平性缺陷不仅可能使受保护群体 [4], [5], [6], [7], [8], 处于不利地位，还可能违反法定要求 [9]。

> 本文介绍了 DICE，这是一个用于深度神经网络公平性缺陷的信息论测试和调试框架。

量化公平性。软件工程和机器学习领域的共同努力，已促成多个成功的公平性测试 框架 [10], [11], [12], [13]。这些框架刻画了各种公平性概念——例如群体公平性 [14]（不同受保护群体的决策结果必须相似）和个体公平性 [15]（仅在受保护属性上存在差异的个体必须获得相似的结果）——并采用基于搜索的测试来发现公平性缺陷。虽然公平性的二元分类有助于发现缺陷，但开发者可能需要进一步了解这些缺陷的性质，以决定如何进行潜在的“漏洞修复”。有些缺陷是否比其他缺陷更严重？这些缺陷是源于训练数据中存在的偏差，还是训练不足造成的假象？能否找到一种不使用受保护信息来解释训练数据的替代方案？

个体歧视是一种经过深入研究的 [16], [11], [17], [18] 因果公平性概念：如果存在另一个个体（可能是反事实个体），它仅在受保护特征上有所不同，却获得了更有利的结果，那么该函数就被定义为对某个个体（输入）存在歧视。我们提出了这一概念的定量推广，即定量个体歧视（QID）。我们将 QID 定义为推导结果时所使用的受保护信息量，并用香农熵、最小熵等熵度量来表征。请注意，QID 度量值为零意味着不存在个体歧视。QID 度量能够按照严重程度对各种具有歧视性的输入进行排序。例如，在某个应用不应依据受保护信息做出决策的情况下，依赖程度更高的输入

则表明违规情况更为严重。我们的第一个研究问题（**RQ1**）关注 QID 度量在查找严重程度不同的输入方面是否有用。

基于搜索的测试。 基于搜索的软件测试提供可扩展的优化算法，以自动发现软件缺陷。在公平性缺陷这一背景下，搜索此类缺陷涉及找到表现出歧视性实例的孪生输入。用于公平性测试的最先进水平算法 [19], [17], [18] 在二值化反馈的驱动下探索输入空间，从而形成不连续搜索域。另一方面，基于 QID 的搜索算法可在优化过程中利用平滑（定量）反馈，从而实现更有指导性的搜索。我们的下一个研究问题（**研究问题 2**）是探究这一理论前景能否在实践中兑现，即能否发现比使用经典歧视概念更丰富的歧视性实例。

因果解释。 尽管按严重程度排序的歧视实例清楚地证明了深度神经网络中存在公平性缺陷，但尚不清楚这些缺陷是训练数据固有的，还是训练过程造成的结果。受“平均因果效应” [20]这一概念以及用于深度学习模型缺陷定位的 AUDEE 框架 [21]启发，我们开发了一种用于公平性缺陷的层与神经元定位框架。如果发现缺陷的原因位于输入层，则表明训练数据中存在歧视。另一方面，如果我们将缺陷原因定位到某个内部层，就希望进一步探查深度神经网络，提取关于神经元及其反事实参数的定量信息，从而在保持准确率的同时缓解缺陷。这项调试工作促使我们提出下一个研究问题（**RQ3**）：是否能够识别一组神经元及其对 QID 的因果效应，以指导缓解措施，同时不影响准确率?

实验。 DICE 实现了一种搜索算法（算法1），用于发现能够最大化 QID的输入；同时实现了一种因果调试算法（算法 2），用于定位在因果上影响 QID数量的层和神经元。我们使用算法公平性文献中的10 个社会关键型深度神经网络，展示了 DICE能够找到在决策中使用大量受保护信息的输入；在生成歧视性实例方面优于三种最先进技术 [19], [17], [18]；并且能够定位神经元，从而指导一种简单的缓解策略，在准确率最多损失 5% 的情况下，将 QID 降低至报告的初始 QID 的 15%。本文的主要贡献如下:

1) 定量个体歧视。 我们提出一种基于香农熵和最小熵的歧视信息论刻画，称为定量个体歧视（QID）。 **2) 基于搜索的测试**。 我们提出一种基于搜索的算法，用于发现深度神经网络表现出严重歧视的情形。 **3) 因果调试**。我们开发了一种因果公平性调试——基于干预语言进行调试，以定位公平性缺陷的根因。

4) 实验评估。在不同数据集和深度神经网络模型上进行的大量实验表明了其可行性、实用性和可扩展性（相较于最先进水平）。我们的框架可以处理多个受保护属性，并且可以轻松适配回归任务。

II. 预备知识

公平性术语。 我们将决策支持系统视为二元分类器，如果预测标签为有利，则表示该预测为输入（个人）带来了理想结果。这些有利预测可能包括对贷款的较高收入估计、假释评估中较低的再次犯罪风险，以及较高的类别不及格风险。每个数据集由若干属性 （例如收入、经历、先前被逮捕记录、性别和种族）以及一组实例 组成，用于描述每个个人的属性值。根据伦理和法律要求，数据驱动的软件不应基于个人的歧视，例如其受保护属性，包括性别、种族、年龄、残障、肤色、信仰、国籍、宗教、遗传信息、婚姻状况和性取向。

存在几种公认的公平性定义。群体公平性要求不同受保护群体 的机器学习结果统计量相似 [14] 使用诸如机会均等差异 （EOD）之类的指标进行衡量，EOD是两个受保护群体的真正例率（TPR）之差。无意识公平（FTU） [15] 要求在训练期间移除受保护属性。然而，FTU可能无法提供充分的支持，因为受保护属性可以通过非受保护碰撞属性（例如种族和邮政编码）影响预测。基于认知的公平（FTA） [15] 是一种个体公平性理念，要求两个个体若根据其非受保护属性被认为相似，则应受到相似的对待。我们的方法致力于实现个体公平性。

个体歧视。 因果歧视最早由 THEMIS [10], 研究，通过反事实 查询来衡量两个子组之间的差异。它对受保护属性设为 A 的个体进行抽样，并将结果与受保护属性设为 B 的反事实情景进行比较。个体歧视（ID）是一种普遍的概念，它通过调整反事实查询，寻找这样一个个体：当其受保护属性不同时，反事实情景会得到更有利的结果。最先进水平的公平性测试使用这一公平性概念来生成公平性缺陷 [22], [19], [17], [18] ，它也与情境测试概念 [23]密切相关。虽然标准的群体公平性指标（例如 AOD/EOD）已有定量定义，但个体公平性尚无定量衡量方法。我们提出采用信息论工具，为个体公平性提供定量衡量方法。

信息论概念。 R´ényi 熵 [24], $H_\alpha(X)$用于量化系统对输入 X作出响应时的不确定性（随机性）。特别是，香农

熵（α=1）和最小熵（α=∞）是 R´enyi 熵的两个重要子类别。香农熵（H_1）衡量有限多个事件的不确定性期望值，而最小熵（H_∞）衡量单个最大似然事件的不确定性。

考虑一个确定性系统（如预训练深度神经网络），其响应集合有限，并假设输入 X 服从均匀分布。因此，该系统在输入集合 X 上诱导出一个等价关系，使得当两个输入的系统输出近似接近时，它们是等价的，即 $x \sim x'$当且仅当 深度神经网络$(x) \approx_\epsilon$深度神经网络 (x')。令 X_o 表示输出为 o 的 X 的等价类。那么，在观察到 DNN 对 X 的输出后，剩余的不确定性可表示为：

$$H_1(X|O) = \sum_{O=o} \frac{|X_o|}{|X|} . \log_2(|X_o|) \qquad \text{（香农熵）}$$

其中，$|X|$ 是 X 的基数，$|X_o|$ 是输出 o的等价类大小。类似地，最小熵定义为

$$H_\infty(X|O) = \log_2(\frac{|X|}{|O|}) \qquad \text{(min-entropy)}$$

其中，$|O|$ 是关于 X [25]、[26], [27] 的等价类数量。鉴于两种熵的初始熵均等于 $\log_2(|X|)$，系统用于做出决策的来自 X 的信息量是

$$\begin{aligned} I_1(X;O) &= \log_2(|X|) - H_1(X|O), \text{ and} \\ I_\infty(X;O) &= \log_2(|O|), \end{aligned}$$

在香农熵和最小熵下，使用 $I_1 \leq I_\infty$。

定量概念与公平性。 我们的方法与这些最先进技术 [28], [22], [19], [17], [29]、[18] 不同之处在于，我们将个体歧视概念扩展为定量信息流，使我们能够衡量基于机器学习的软件系统中的歧视程度。给定非受保护属性和机器学习结果，香农熵 衡量受保护类别变化时所有可能响应中个体歧视量的期望值，而最小熵 则衡量来自最大似然类别的单个响应中的歧视量。

示例 1. 考虑一个数据集，其中有 16 种不同的受保护值——性别（*2*）、种族（*2*）和年龄（*4*），且分布均匀；假设系统中有 4 名个体。我们扰动这些个体的受保护属性，生成16个反事实样本，并将其输入深度神经网络以获取预测分数。假设对于第一名个体，所有16个输出都属于同一个类别（绝对公平）；对于第二名个体，则有 16 个类别，每个类别各有一个输出（绝对歧视）。对于第三名个体，设输出属于 4 个类别，每个类别有 $\{4, 4, 4, 4\}$ 个元素（例如，每个年龄组对应一个输出类别）。对于第四名个体，设输出属于 5 个类别，每个类别有 $\{8, 4, 2, 1, 1\}$ 个元素（例如，如果种族=*1*，则输出为类别 1；否则，如果性别=*1*，则输出娱乐主管是2；否则如果年龄=$\{1,2\}$，则输出为3；否则每个年龄=$\{3,4\}$对应一个输出类别）。

我们采用以下歧视概念。

- 个体歧视概念。最先进技术所采用的个体歧视概念只能区分第一个个体与其余个体，却无法区分第二至第四个个体。事实上，这些技术能在短时间内生成数万个个体歧视实例 [19], [17], [18]。然而，它们无法优先安排用于缓解的测试用例，也无法刻画歧视程度（即其严重性）。

- 香农熵。使用香农熵，我们得出初始公平性为 4。0 比特，即最大可能的歧视。对于第一至第四位个体，深度神经网络剩余的公平性分别为 4。0、0。0、2。0和 2。125。歧视是初始公平性与剩余公平性之间的差值，对于第一至第四位个体，分别为 0。0、4。0、2。0 和 1。875。需要注意的是，除两个极端情况外，香农熵认为，对第三位个体（而非第四位个体）的扰动会产生更高程度的歧视。

- 最小熵。基于最小熵的初始公平性也是 4 比特。对于这四个人，条件最小熵分别为 $\log \frac{16}{1} = 4.0$、$\log \frac{16}{16} = \overline{0}.0$、$\log \frac{16}{4} = \overline{2}.0$ 和 $\log \frac{16}{5} = \overline{1}.7$。因此，歧视程度分别为 0.0、4.0、$\log 4 = 2.0$ 和 $\log 5 = 2.3$。除两个熵相等的极端情况外，最小熵认为，对第四个人的扰动会造成更高程度的歧视。这很直观，因为第四种情况中的歧视更加微妙、复杂且显著。因此，机器学习软件开发人员或许会优先处理由最小熵刻画的那些情况。

III. 概述

DICE 简述。 图 1 展示了我们提出的框架 DICE 的概览。该框架由两个组件组成：（1）基于搜索算法的自动测试生成机制；（2）一种调试方法，利用因果算法定位对公平性有显著影响的神经元。首先，DICE 在输入数据集空间中进行搜索，以找到这样的非受保护属性条件：在这些条件下，被测深度神经网络在做出决策时明显依赖受保护属性。为此，它分为全局阶段和局部阶段。在全局阶段，搜索逐步探索输入空间，以增加每一步搜索中的歧视程度。另一方面，在局部阶段，它利用全局阶段中有希望的种子，尽可能生成更多歧视性实例。

搜索的关键要素是一种基于阈值的聚类算法，用于计算梯度和目标函数，从而提供平滑的反馈。搜索过程刻画了定量个体歧视（QID）和

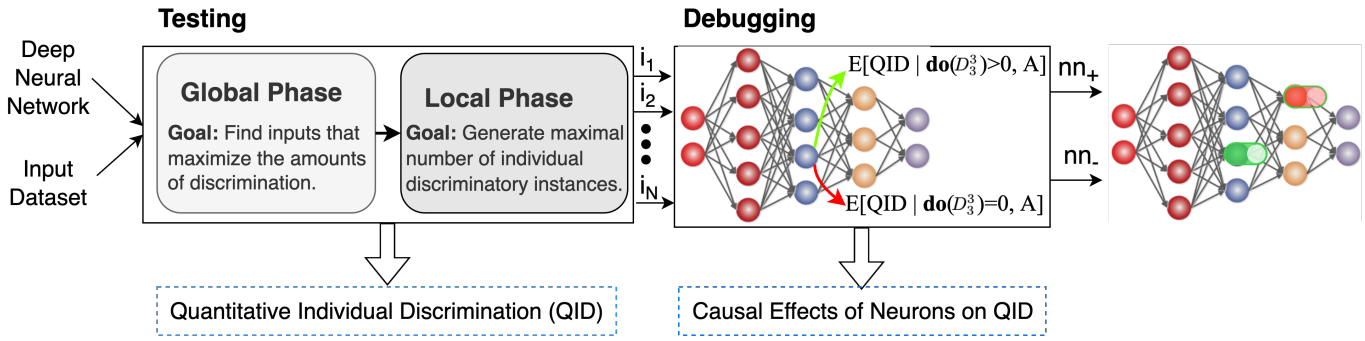


图 1：DICE 的工作流程。给定一个深度神经网络和相关输入数据集，DICE 通过测试量化 QID 歧视，并应用因果调试来定位和缓解 QID 歧视。

返回一组有趣的输入。其次，DICE 使用这些输入来定位对歧视程度具有最大因果效应的神经元。在此过程中，它会对一组可疑神经元进行干预 [30] 。对于每个神经元，我们的调试方法会在测试用例上强制神经元处于激活状态（n>0）和非激活状态（n=0），前提是 DNN 的功能准确率保持在有效范围内。然后，它计算这两种情况下 QID 程度之间的差异，以刻画该神经元对公平性的因果效应。

DICE报告排名靠前的 k 个神经元，这些神经元既具有正面影响（即其激活会减少歧视程度），也具有负面影响（即其激活会增加歧视程度）。一种潜在的缓解策略是进行干预，使一小组神经元保持激活状态（对于正神经元）或保持停用状态（对于负神经元）。

测试用例。考虑使用预训练模型对具有 6 层 [17]的成人普查收入 [31] 数据集，以概述 DICE 的实际应用。我们运行 DICE 1 小时，获得了 230,593 个测试用例。它从最初的 14 个聚类簇中发现了 36 个聚类簇，QID 的数量分别为 4.05 位和 2.64 位的最小熵和香农熵，而受保护属性的信息总量为 5.3 位。按测试用例排序后，我们有 6 个测试用例具有最大 QID 歧视，为 5.3 位。此外，还有 29 个和 112 个测试用例，其 QID 歧视分别为 5.2 位和 5.1 位。报告的数值为 10 次运行的平均值（并已四舍五入）。

定位与缓解。 DICE 使用生成的测试用例定位对歧视有显著因果贡献的层和神经元。对于人口普查数据集，它识别出第二层是对受保护属性最敏感的层。在该层的神经元中，DICE 发现第 15 个神经元对公平性有最大的负面影响（停用该神经元后，歧视减少了 19.6%），而第 19 个神经元对公平性有最大的正面影响（激活该神经元后，歧视减少了 17.6%）。基于这一定位，采用激活或停用这些神经元这一简单的缓解策略，可将 QID 歧视程度降低 20%，准确率损失为 3%。

与最先进水平的比较。 我们从以下方面将 DICE 与最先进技术进行比较：针对每个受保护属性生成个体歧视（ID）（而非定量意义上的歧视）。我们的目标是评估基于聚类的搜索在生成歧视性实例方面是否有效。我们运行 DICE 和基线方法 15 分钟，并报告 10 次运行的平均结果。基线方法包括 AEQUITAS [19], ADF [17],和 NEURONFAIR [18]。以人口普查数据集中的性别作为受保护属性时，DICE 生成了 $79.0k$ 个实例，而 AEQUITAS、ADF 和 NEURONFAIR 分别生成了 $10.4k$、$18.2k$ 和 $21.6k$ 个歧视性实例。总体而言，在所有情况下，DICE 都生成了更多 ID 实例，并且成功率更高。然而，DICE 找到第一个 ID 实例的速度较慢，平均需要几秒钟，因为我们的方法在全局阶段不会生成 ID 实例。当考虑找到前 1,000 个实例所需的时间时，DICE 显著优于最先进水平。我们推测，这些改进得益于通过定量反馈使搜索空间更加平滑。

IV. 问题陈述

我们考虑输入变量集合为 A 的基于深度神经网络的分类器，其中输入变量被划分为受保护变量集合 Z（例如种族、性别和年龄）和非受保护变量集合 X（例如职业、收入和教育程度）。我们进一步假设输出由 t 个预测类别组成。

定义IV.1 (DNN：语义)。深度神经网络（深度神经网络）编码了一个函数 $\mathcal{D}$： $X \times Z \rightarrow [0,1]^t$ 其中 $X = X_1 \times X_2 \cdots \times X_n$ 是非受保护输入变量的集合，$Z = Z_1 \times Z_2 \cdots \times Z_r$ 是受保护变量的集合，输出是一个 t 维概率向量，对应于 t 个预测类别。输入对 (x, z) 的预测标签$\mathcal{D}_\ell(x, z)$ 是得分最高的索引，即 $\mathcal{D}_\ell(x, z)$ $=$最大值$_i \mathcal{D}_\ell(x, z)$ (i)。我们假设受保护输入变量集合的定义域是有限的，并令 m为受保护变量集合 Z的基数。

定义 IV.2 (深度神经网络：语法)。 一个深度神经网络 $\mathcal{D}$ 由输入维度 n+r、输出维度 t、隐藏层的深度 N以及其隐藏层的权重$W_1, W_2, \ldots, W_N$进行参数化。我们的目标是测试并调试一个参数和权重已知的预训练神经网络。令 D_i

是层 i 的输出，该层实现了从前一层的输出 D_{i-1} 及其权重 W_{i-1} 到 $1 \leq i \leq N$ 的仿射映射，随后

1) 一个固定的非线性激活单元（例如，定义为$D_{i-1} \mapsto \max\{W_{i-1}.D_{i-1}, 0\}$的 *ReLU*），用于 $1 \leq i < N$；或者 *2)* 一个 *SoftMax* 函数，将分数映射为 $i = N$中每个类别的概率。

令 D_i^j 为第 i 层神经元 j 的输出。

个体歧视。 如果深度神经网络 $\mathcal{D}$ 基于因果歧视概念 [16], [11], [17], [18] 存在偏差，我们称其

$$\exists z_1, z_2 \in Z, x \in X \ s.t. \ \mathcal{D}_\ell(x, z_1) \neq \mathcal{D}_\ell(x, z_2),$$

对于 $z_1 \neq z_2$ 个受保护输入。直观而言，其思路是找到一个个体，使其在受保护属性（例如种族）不同的反事实情况下得到不同的结果。

定量个体歧视。 在公平性测试中，通常需要量化个体所受偏差的程度。我们基于深度神经网络在受保护属性上的输出所诱导的等价类，定义定量个体歧视（QID）这一概念。形式上，$QID(Z, X = x) = \langle Z_1, \ldots, Z_k \rangle$，即由深度神经网络输出刻画的 Z 商空间，其中个体具有非受保护值 x。利用这一概念，若一对受保护值 z, z' 属于同一等价类 i（即 $z, z' \in Z_i$），当且仅当 $\mathcal{D}(z, x) \approx \mathcal{D}(z', x)$，我们就称它们属于同一等价类。

鉴于 Z 服从均匀分布，且 $\mathcal{D}$ 是确定性函数，我们可以分别根据香农熵和最小熵来量化个体 (z, x) 的 QID 概念：

$$Q_1(Z, x) = \log_2(m) - \sum_{i=1}^{k} \frac{|QID_i(Z, x)|}{m} . \log_2(|QID_i(Z, x)|)$$

$$Q_\infty(Z, x) = \log_2(m) - \log_2(\frac{m}{k}) = \log_2(k).$$

其中，m 是 Z的基数，$|QID_i(Z, x)|$ 是等价类 i的大小，k 是等价类的数量。

针对 QID 调试/缓解深度神经网络。通过QID刻画歧视程度后，下一步是定位一组会因果性地影响深度神经网络输出、使其具有 k 个等价类的层和神经元。

因果逻辑 [30] 为推理变量间的因果关系提供了坚实基础。我们考虑一个结构因果模型（SCM），其中外生变量U表示未观测到的输入因素，内生变量V表示 (X, Z, D_i^j)；函数集合 $\mathcal{F}$定义在集合V 上，并使用深度神经网络函数 $\mathcal{D}$ 和外生变量 U。借助 SCM，我们旨在估计神经元 D_i^j 对 QID 的平均因果效应（ACE） [30] 。

执行此类计算的主要工具称为 do 逻辑 [20]。我们用 $do(i, j, y)$ 表示对第 i 层的神经元 j 进行干预，使其输出保持为 y。为此，我们移除指向该神经元的入边，并强制其输出取预先定义的值 y，但我们由于深度神经网络的前馈结构，无需控制后门变量。接着，在最小熵条件下，神经元 D_i^j 对定量个体歧视的因果效应（ACE）可写为 $E[Q_\infty | do(i, j, y), k, l]$，它表示在非干预深度神经网络将 k 个类别分类且准确率为 l 的条件下，对该神经元进行干预后所得到的预期 QID。我们的目标是找出对 QID 具有最大因果效应的神经元，同时要求此类干预忠实于深度神经网络的功能。

定义 IV.3（定量公平性测试与调试）。 给定一个在数据集 A 上训练的深度神经网络模型 $\mathcal{D}$，其属性包括受保护属性（$Z \subset A$）和非受保护属性（$X \subset A$）；搜索问题是，对于选定的度量 Q_1 或 Q_∞，在 m 个受保护值 $\Sigma = \{z_1, \ldots, z_m\}$中，找到一个使定量个体歧视（*QID*）最大化的单个非受保护值 $x \in X$。给定表征最大 *QID* 的输入（Σ, x），我们的调试问题是为 $J \subseteq |W_l|$ 找到一个最小层子集 $l \subset \{1, \ldots, N\}$ 和神经元子集 $D_l^{(J)}$，使 D_l^J 对 *QID* 的平均因果效应最大。

V. 方法

表征定量个体歧视。给定数据集 A上的深度神经网络 $\mathcal{D}$，我们的目标是表征所有可能个体中的最坏情况 QID。由于最小熵刻画了来自一次预测的歧视程度，其中预测结果为 Q_∞（Z，x）和 $\geq Q_1$（Z，x），因此它是优先安排测试用例的有用概念。因此，我们聚焦于 Q_∞，并提出以下目标函数：

$$\max_{x \in X} \ 2^{Q_\infty(Z, x)} + (1 - exp(-0.1 * \delta))$$

其中 $2^{Q_\infty(Z, x)} = k$ 是等价类之间的最大距离，并通过指数函数进行归一化，使其取值保持在 0 到 1 之间。当两个实例所表征的类别数量相同时，该项用于打破平局，优先选择距离最大的实例。总的来说，目标是找到非受保护属性 x 的单一取值，使神经网络模型 $\mathcal{D}$ 在将 x 与 m 受保护值配对时，能够预测出许多可区分的结果类别。然而，找到这些输入需要在输入空间的指数级子集集合中进行穷举搜索，因此显然难以实现。我们提出一种梯度引导的搜索算法，旨在搜索输入变量（属性）的空间，以最大化等价类的数量，并尽可能生成更多歧视性实例。

搜索方法。 我们的搜索策略与一些先前工作 [17], [29], [18]类似，由全局阶段和局部阶段组成。全局阶段旨在通过梯度引导的聚类，找到最大的定量个体歧视。局部阶段则利用有希望的实例，生成尽可能多的歧视性实例（ID）。

全局阶段。 给定当前实例 x，全局阶段首先使用受保护属性空间中的 m 个不同值，同时保持非受保护属性的值不变。然后，它从深度神经网络接收 m 个预测分数，并

将它们划分为 k 类别。我们采用一种基于约束的聚类方法，其中 ϵ ：如果两个元素的得分相差超过 ϵ，它们就不能属于同一个聚类簇。现在，关键步骤是沿着一个方向，对当前实例的一部分非受保护属性进行扰动；该方向很可能会增加下一步全局搜索中由扰动后的实例产生的聚类簇数量。

为此，我们首先计算最大聚类簇中的一对实例（例如 a， a'）的深度神经网络损失函数梯度。其直觉是，我们更有可能将最大的聚类簇拆分为 2 个或更多子聚类簇，并在下一步增加划分数量。对于这对样本，我们使用梯度方向相同的非受保护属性 d ，因为这表明损失函数对这对样本共有特征上的微小变化具有较高的敏感性。如果使用方向相反的梯度，由于我们只扰动非受保护属性上的一个实例，梯度的影响就会相互抵消。最后，我们扰动当前样本 x，使用方向 d 和步长 s_g 生成 x' 。

局部阶段。 一旦检测到某个实例具有超过 2 个聚类簇，我们就进入局部阶段，目标是尽可能生成更多歧视性实例(ID)。在我们的定量方法中，如果存在一个获得了有利结果的反事实个体 x' ，那么我们就称针对个体 x 作出的不利决策具有歧视性。与最先进水平 [19], [17], [18], 类似，我们使用非线性优化器，它以一个初始实例 x、一个用于在当前实例的邻域中生成下一个实例的步进函数，以及一个用于量化当前实例所受歧视程度的目标函数作为输入。由于我们的方法采用基于聚类簇特征的连续目标函数，因此能够引导局部搜索生成歧视性实例。

搜索过程。 算法1概述了我们用于量化偏差程度的搜索算法。我们首先使用聚类（KMeans算法）将数据点划分为 p 组（第1行）。接着，我们运行该算法，直到达到超时 T；每次迭代中，我们从某个划分中随机选取一个样本作为种子 p（第2行）。然后，我们进入搜索的全局阶段和局部阶段。

在全局阶段，我们首先使用种子实例 x ，并扰动受保护属性 P ，以生成 m 个可能的实例 X_m ，这些实例具有不同的受保护值，而非受保护值相同。然后，我们将 X_m 输入DNN模型，以获取概率分数 S_m （第5行）。接着，我们使用容差 ϵ 将这些分数 S_m 聚类为 k 组（第6行）。之后，我们计算两个随机实例（取自规模最大的聚类簇）的梯度，并选取在梯度方向上达成一致次数最多的一部分非受保护特征。然后，我们使用这些特征及其方向扰动输入 x ，生成下一个实例 x' （第7–10行）。如果聚类簇数量或簇间距离增加，则进入局部阶段（第11行）。

在局部阶段，我们使用通用优化器，

算法1： DICE（搜索）

输入： 数据集 A、深度学习模型 $\mathcal{D}$、深度神经网络的损失函数 J、受保护属性 P、非受保护属性 NP、数据集的划分数量 p、全局扰动的步长 s_g、局部扰动的步长 s_l、全局迭代的最大次数 N_g、局部迭代的最大次数 N_l、容差 ϵ 和超时 T。

输出： 聚类簇数量和（局部+全局）测试用例。

```
1  A', cur ← KMeans(A, p), time()
2  当 time()- cur < T 时执行
3      x, i, k, δ ← pick(A'), 0, 1, 0.0
4      当 i < N_g do
5          I_m, S_m ← 生成_预测(a, P)
6          X_k, δ' ← Clust(S_m, ε)
7          a, a' ← 选择_配对_最大(X_k)
8          Gs ← (∇J(a) ∇J(a')
9          d ← 选择_常见_直接(Gs, NP)
10         x' ← 扰动(x, d, s_g)
11         如果 (|X_k| > k) 或 (|X_k| = k and δ' > δ) 那么
12             eval_f(x) {
13              I'_m, S'_m ← 生成_预测(x, P)
14              X_k' ← Clust(S'_m, ε)
15              Δ ← arg.max(X_k')−arg.min(X_k')
16              local_inps.add(x)
17              返回 −Δ}
18             步骤_f ← λ_x 扰动_局部(x, s_l)
19             LBFGS(x, 评估_f, 步骤_f, N_l)
20         global_inps.添加(a)
21         k, x ← 最大值(k, |X_k|), x'
22 返回 k, I = global inps ∪ local inps
```

称为 LBFGS [32]，它接收一个初始种子 x、一个目标函数、一个步进函数以及局部迭代的最大次数 N_l；它返回优化过程中生成的实例（第 12—19 行）。在 eval_f 所示的目标函数中（第 12—17 行），我们生成 m 个实例，这些实例具有相同的非受保护值，但受保护值不同（第 13 行）。我们为这些实例生成预测分数，并使用容差参数 ϵ 对其进行聚类（第 14 行）。然后，我们计算得分最小和最大的两个聚类簇的索引之差（第 15 行）。最后，我们记录生成样本，并将该差值作为优化器在当前样本处的评估值返回（第 16—17 行）。步进函数如 perturb_local 所示（第 18 行），它引导优化器在输入空间中执行一步。我们的步进函数从与当前样本不同的聚类簇中随机采样。然后，它计算归一化梯度和，并使用最小的梯度对其进行扰动，使其保持在当前样本的邻域内。

调试方法。 由于在深度神经网络中对所有可能的神经元进行干预在计算上十分困难，我们首先

借鉴深度学习框架调试文献中的层定位技术 [33], [21], 我们检测对受保护属性最敏感的层。令 $D_i(z,x)$ 为层 i 在受保护值 z 和非受保护值 x下的输出。令 $\Delta_i(x)$: $\mathbb{R}^{|D_i|} \times \mathbb{R}^{|D_i|} \to \mathbb{R}$ 表示由 m 个不同的受保护值和相同的非受保护值 x触发的深度神经网络在层 i 处的输出之间的距离，并令 δ_i 为 $\max_x \Delta_i(x)$。层 i （关于受保护属性）的敏感性变化率为

$$\rho_i = \frac{\delta_i - \max_j \delta_j}{\max_j \delta_j + \epsilon},\ with\ 0 \leq\ j <\ i$$

其中 $\delta_0 = 0.0$ 和 $\epsilon = 10^{-7}$ （为避免除以零 [33], [21]）。令 $l = \arg\max_i \rho_i$ 为变化率最大的层索引。下一步，我们将定位层 l 中对公平性有显著正向或负向影响的神经元。令 V_l^j 为神经元 j 在后续 l 时可能取值的集合（在层定位期间记录）。我们希望计算神经元 D_l^j 激活与未激活时的平均因果效应，同时注意到此类干预可能会影响深度神经网络的功能。因此，在一组干预值中，我们选择一个激活值 $v_1 \in V_l^j > 0$ 和一个未激活值 $v_2 \in V_l^j \approx 0$，使深度神经网络的功能准确率保持在原始准确率 A 的 ϵ 以内。因此，我们将神经元 D_l^j 的平均因果差异（ACD）定义为:

$$\mathbb{E}[Q_\infty \mid \mathrm{do}(l,j,v_1>0),k,A] - \mathbb{E}[Q_\infty \mid \mathrm{do}(l,j,v_2\approx 0),k,A],$$

其中，do 算子用于强制将层 l 中神经元j 的输出设为固定值 v。然后，我们返回正值最大（加剧歧视）^^和负值最小（缓解歧视）的神经元索引。令 i 和 j 分别表示 ACD 最大的层和神经元。因此，一种简单的缓解策略是停用神经元D_l^j，预计可将QID降低 ACD/k 个百分点。同样，激活负值最小的神经元 ACD 预计可将 QID 降低 ACD/k 个百分点。

调试流程。算法2 展示了 DICE的调试方面。给定一组来自搜索算法的测试用例，我们首先使用距离的概念（例如，$\Delta = L_1$）来计算任意一对受保护值z, z' 相对于层 $l \in \{1,\ldots,N\}$ 的输出之间的差异（第 1行）。然后，我们计算变化率（第 2-3行），并返回变化最大的层 l（第 4 行）。我们计算每个神经元 i 在层 l 输出上的各种统计量，例如最小值、最大值、平均值、平均值±标准差、平均值±2*标准差等（第5 行）。在这些值中，我们取最小值和最大值，使得对层 l 上的神经元 i 进行干预时，对深度神经网络的准确率影响最小（第6行）。最后，我们计算平均因果差异（第7 行），并返回层的索引、具有较大负面影响的神经元以及具有较大正面影响的神经元。

算法 2： DICE（调试）。

输入： 数据集 $A = (A_X,\ A_Z)$, $\mathcal{D}$ 准确率为 $\mathcal{A}_\mathcal{D}$, 测试用例 I, 距离函数 Δ , the 层定位的容差 ϵ_1，该容差 的准确率损失 ϵ_2，以及 k 个顶部项目。

输出： 层索引、负神经元和正神经元。

/* 层定位 */

$\delta \leftarrow \lambda_l$最大化 $\sum_{x\in I} \Delta\big(D_l(z,x)\ D_l(z',x)\big)$

$\delta[0]$, $\delta_{max} \leftarrow 0.0$, λ_i 最大$_{j<i}\ \delta[j]$

$\rho \leftarrow \lambda_l \frac{\delta[i]-\delta_{max}[i]}{\delta_{max}[i]+\epsilon_1}$

$l \leftarrow \arg\max_i \rho[i]$

/* 神经元定位 */

$V_l \leftarrow \lambda_i$ 统计信息(D_l^i)

$v_l \leftarrow \lambda_i\ \lambda_j\ |\mathcal{A}_\mathcal{D} - \mathcal{A}_{\mathcal{D}\leftarrow do(l,i,V_l[j])}| \leq \epsilon_2$

$ACD_l \leftarrow \lambda_j\ \mathbb{E}\,(k'|\mathrm{do}(l,j,v_l^j>0))$ - $\mathbb{E}(k'|\mathrm{do}(l,j,v_l^j=0))$

返回 l, top_k(最大 ACD_l), top_k(最小 ACD_l).

VI. 实验

数据集和DNN模型。我们考虑了算法公平性文献中的 10 个具有重要社会意义的数据集。这些数据集及其属性见表 I。对于 DNN 模型，我们采用了文献 [17], [18], [29]中的相同架构，并使用包含 $\langle 64,32,16,8,4,2\rangle$ 个神经元的六层全连接神经网络训练所有数据集。所有训练均使用相同的超参数：训练轮数、批量大小和学习率分别设为 1000、128 和 0。01。训练所得模型的准确率见表 V。

技术细节。 我们使用 TensorFlowv2.7.0和 scikit-learn v0.22.2实现了 DICE。我们在配备 AMDRyzen Threadripper PRO 3955WX 3.9GHz 16 核X 32 CPU 和两块 NVIDIA GeForce RTX3090 GPU的 Ubuntu 20.04.4LTS操作系统服务器上运行了所有实验。对于算法1中的最大全局值、最大局部值、 ϵ、 $s\ g$ 和$s\ l$，我们分别取10、1000、0.025、1和1，并对所有实验进行 10次重复运行后取平均值。在算法 2 中，我们分别将 L_1-范数、 10^{-7}、0.05和 3 用作 Δ、 ϵ_1、 ϵ_2 和 k。

研究问题。 我们通过实验设置寻求回答以下三个问题。

RQ1DICE 能否刻画推理过程中使用的受保护属性信息量? **RQ2**所提出的搜索算法在生成个体歧视实例方面是否有效且高效（相较于最先进技术）？ **RQ3** 所提出的因果调试能否指导我们定位并缓解歧视程度? 我们开源的工具 DICE 及所有实验对象均可公开访问:

- https://github.com/armanunix/Fairness-testing

A. 通过搜索刻画 *QID*（*RQ1*）

一个重要目标是刻画在推理 DNN模型时使用了多少来自受保护属性的信息。表 II 展示了回答这一研究问题的实验结果。表格左侧显示了

表 I：我们实验中使用的数据集。

数据集	#实例	#特征	受保护群体 *Name*	*Size*	数量。受保护值 (m)	结果标签 标签 *1*	标签 *0*
成人 人口普查 收入 [31]	32,561	13	Sex	2	90	高收入	低收入
			Race	5			
			Age	9			
Compas [34]	7，214	12	Sex	2	12	未再犯罪	再犯
			Race	2			
			Age	3			
德通 信用 [35]	600	20	Sex	2	16	良好信用	不良信用
			Age	8			
违约 信用 [36]	13,636	23	Sex	2	12	违约	未违约
			Age	6			
心脏 健康 [37]	297	13	Sex	2	14	疾病	非疾病
			Age	7			
银行 营销 [38]	45，211	16	Age	9	9	订阅者	非订阅者
糖尿病 [39]	768	8	Age	9	9	阳性	阴性
学生 表现 [40]	1044	32	Sex	2	16	Pass	未通过
			Age	8			
MEPS15 [41]	15,830	137	Age	9	36	所利用的益处	未充分利用的福利
			Race	2			
			Sex	2			
MEPS16 [41]	15,675	137	Age	9	36	已利用的效益	未使用的福利
			Race	2			
			Sex	2			

初始特征，例如受保护值的数量（m）、根据 $\min(\epsilon^{-1}, m)$ 计算的最大可能歧视量（Q_I），以及使用数据集样本发现的初始聚类簇数量（K_I）。表格右侧显示了运行搜索 1 小时后的结果。列 $\#I$ 表示生成的 QID 实例数量，K_F 表示 DICE 发现的最大簇数。列 $\mathrm{T}K_F$ 表示从初始簇数为 K_I 的输入中找到最大簇数所用的时间（秒）。列 Q_∞ 和 Q_1 分别表示基于最小熵和香农熵计算的定量个体歧视。列 $\#I_{K_F^1}$、$\#I_{K_F^2}$ 和 $\#I_{K_F^3}$ 分别显示具有最高、次高和第三高 QID 的测试用例数量，这些 QID 按严重程度对测试用例排序。总体而言，结果表明，DICE 可以在一分钟内搜索，从初始特征中找到更多的簇（平均而言）。学生 DNN 数据集的簇数增幅最大，从 1。9 增至 10。9。DICE 发现，成人收入普查数据集的 QID 数量最多：受保护变量中的 5。3 位中有 4。05 位（76。4%）被用于决策。德国信贷数据集的歧视量最少，为 4。0 位中的 1。61 位（40。0%）。对于测试用例优先级排序，列 $\#I_{K_F^1}$ 表明，我们的方法有助于从生成的测试用例中找出一小部分具有最严重歧视的用例。在 10 项实验中，有 7 项中，DICE 从数十万个输入中找出的严重歧视测试用例少于 50 个。

回答研究问题1： 该搜索算法能够有效地通过 QID刻画歧视程度。在1 小时内，它平均将聚类簇数量增加了 $3.4\times$，并找到了使用高达 76%的受保护信息（5.3 比特中的4.05 比特）来推断DNN输出的实例。DICE 有助于根据严重程度确定测试用例的优先级；在数十万个测试用例中，它生成的具有最大 QID 的测试用例少于 50 个。

B. Individual Discriminatory Instances (RQ2)

在本节中，我们将我们的搜索算法在搜索个体歧视（ID）实例（如第 IV节所定义）方面的效率和有效性，与最先进技术进行了比较。我们的基线方法是 AEQUITAS [19], ADF [17],和NEURONFAIR [18]。我们从这些工具的 GitHub 代码库中获取了其实现，并根据规定的设置进行了配置，以获得最佳性能。遵循这些技术，我们分别报告每个受保护属性的结果。表III展示了基线方法和DICE在运行15 分钟后的结果。结果取10次重复运行的平均值。列$\#ID$ 是生成的个体歧视实例总数。列 l_s 是搜索局部阶段的成功率。我们不计入全局成功率，因为在我们的搜索中，全局阶段的目标是最大化 QID，而局部阶段则侧重于生成许多 ID 实例。我们将成功率计算为找到的 ID 数量除以生成样本总数。列 $T.1st$ 和 $T.1k$分别表示找到第一个ID 实例以及生成1,000 个个体歧视实例所需的时间（以秒计）（注意： N/A 在列 $T.1k$ 中表示该工具在 900 秒的实验超时时间内未生成$1,000$ IDs ，该数值为 10 次运行的平均值）。

结果表明，DICE 在生成大量 ID 实例方面优于当前最先进的方法。具体而言，与 AEQUITAS、ADF 和 NEURONFAIR 相比，DICE 在最佳情况下分别多找到 $27.1\times$、$16.0\times$和 $16.0\times$ 个 ID。与 AEQUITAS、ADF 和 NEURONFAIR 相比，DICE 在最差情况下也分别多生成 $3.2\times$、 $2.3\times$和 $2.6\times$ 个 ID。AEQUITAS、ADF、NEURONFAIR 和 DICE 的局部搜索平均成功率分别为 20.6 %、33.0%、29.6% 和 78.2%。在寻找第一个 ID 所需的时间方面，AEQUITAS 表现最佳，平均耗时为 0.03 秒；而 DICE 平均需要 1.46 秒才能找到第一个 ID。平均而言，DICE 生成 1000 IDs 所需时间最短，为 57.2 秒，而 ADF

表 II：DICE 的特征 针对 10 个数据集和深度神经网络在 1 小时运行中进行表征 QID （结果为 10 次运行的平均值）.

数据集	m	Q_I	K_I	$\#I$	K_F	T_{K_F}	Q_∞	Q_1	$\#I_{K_F^1}$	$\#I_{K_F^2}$	$\#I_{K_F^3}$
人口普查	90	5.3	13.54	230,593	35.61	21.04	4.05	2.64	6.0	28.6	111.6
Compas	12	3.6	3.12	157，968	10.24	6.50	1.81	1.40	35.2	338.7	1，016。9
德语	16	4.0	2.34	245,915	9.56	13.14	1.61	1.10	6.6	16.2	54.8
违约	12	3.6	5.58	258,105	11.26	10.94	2.10	1.78	3,528.8	9,847.2	9,771.0
心脏	14	3.8	4.54	270,029	10.01	11.88	2.31	1.80	21.7	135.2	579.7
Bank	9	3.2	1.45	172,686	8.93	3.68	2.25	1.98	5,118.5	13,513.3	20,438
糖尿病	10	3.3	2.39	504,414	7.90	0.016	1.40	1.11	89.7	609.6	2,310.1
学生	16	4	1.90	133,221	10.90	14	1.93	1.35	16.0	130.7	128.7
MEPS15	36	5.2	7.03	19,673	18.52	31.62	2.61	1.62	2.6	3.5	6.0
MEPS16	36	5.2	9.06	14,266	19.25	49.16	2.21	1.52	2.0	3.5	6.0

耗时 179.1（秒），NEURONFAIR 耗时 135.4（秒），AEQUITAS 耗时最长，为 197.7（秒）。总体而言，我们的实验表明，与三种最先进技术相比，DICE 在生成 ID 实例方面更有效，这在很大程度上得益于局部搜索过程中反馈的平滑性。

回答 RQ2：我们的实验表明，DICE 的表现优于最先进水平的公平性测试技术 [19]，[17]，[18]。在最佳情况下，我们的方法比这些技术多发现了 $20\times$ 个个体歧视（ID）实例，平均成功率几乎高出 $3\times$ 。然而，我们发现，DICE 找到第一个 ID 实例的速度比这些技术慢，耗时约为几秒。

C. 面向公平性的深度神经网络因果调试（*RQ3*）

我们对深度神经网络模型进行实验，以研究所提出的因果调试方法是否有助于识别会显著影响由 QID 表征的歧视程度的层和神经元。表 IV 展示了实验结果（取 10次独立运行的平均值）。前两列展示了定位到的层及其影响（即算法2 中的 l 和 ρ ）。接下来的六列展示了对公平性具有正面影响的排名前 3 的神经元（即激活这些神经元会减少基于 Q_∞ 的歧视程度）。最后六列展示了对公平性具有负面影响的排名前3 的神经元（即激活这些神经元会增加基于Q_∞的歧视程度）。与其他层相比，层索引2 更常被定位，其中层3、4和 5各被定位一次。总体而言，对于具有正面公平性影响的神经元，平均因果差异（ACD）范围为 4% 至 55%；对于具有负面公平性影响的神经元，范围为 0.6% 至 18.3%。

在定位的指导下，DICE 通过干预来激活具有正面公平性影响的神经元，或停用具有负面影响的神经元。表 V 展示了这种缓解策略的结果。列 A 和 K 分别显示 DICE 在缓解前对 DNN 模型报告的准确率和簇数量（对一组随机测试用例取平均值）。列 $A^{=0}$ 和 $K^{=0}$ 分别显示通过停用 负面公平性影响最大的神经元（如表 IV 中神经元 $\bar{1}$ 所示）缓解 DNN 模型后报告的准确率和簇数量。同样，

列 $A^{>0}$ 和 $K^{>0}$ 分别是准确率，以及通过激活公平性影响最高的正向神经元（如表 IV 中的神经元1^+所建议）对DNN模型进行缓解后报告的簇数量。结果表明，激活干预可以在准确率损失3% 的情况下将 QID 歧视降低至少 5%，在准确率损失 2% 的情况下最多降低 64.3%。另一方面，取消激活可以在准确率损失 1%的情况下将公平性提高至少 6%，在损失 2% 的情况下最多提高 27%。

回答 RQ3：DICE 中实现的调试方法识别出了对公平性至少具有 5%、最高达 55% 正因果效应的神经元，以及至少具有 0.6%、最高达 18.3% 负因果效应的神经元。采用缓解策略并进行定位后，可将歧视程度降低至少 6%、最高达 64.3%，且准确率损失低于 5%。

VII. 讨论

限制。在这项工作中，我们考虑所有受保护值的集合，并对其进行扰动以生成反事实样本。对受保护属性进行的各种扰动可能产生不现实的反事实样本，并导致假阳性（对歧视的过度估计）。通过提供特定领域的约束（年龄< YY $\Longrightarrow$NOT(已婚)），可以缓解这一限制：我们已经应用了一些常识性约束（例如，确保年龄处于有效范围内）。此外，与任何动态测试方法类似，我们的方法可能会遗漏具有歧视性的输入，并且容易出现假阴性。在适当的统计检验（例如贝叶斯因子）下，可以控制遗漏相关输入的概率。此外，我们的调试方法类似于定位可疑代码片段，并且基于其对决策影响的因果推理，而不是相关性。但是，它无法为黑盒深度神经网络函数提供解释或诠释。

效度威胁。为解决内部效度问题并确保我们的发现不会导致无效结论，我们遵循既定指南，并对重复实验取平均值。为确保我们的结果具有可推广性并解决外部效度问题，我们在从公平性测试相关文献中选取的10个深度神经网络模型上进行实验。然而，

表 III：在900秒内将生成歧视性实例的方法与最先进水平进行比较。PROT.表示受保护属性；#ID表示个体歧视实例的数量； $l\ s$ 表示局部搜索的成功率； $T.1st$ 和 $T.1k$ 分别表示找到第一个和第$1,000^{th}$ 个歧视性实例所用的时间（单位：秒）。结果取每种工具运行 10 次的平均值，其中括号内报告了相对于均值的偏差。最佳结果以**加粗** 突出显示（k 为 1,000， $\epsilon < 0.1$ 用于计算标准差）。

数据集	协议。	AEQUITAS [19]				ADF [17]				NEURONFAIR [18]				DICE			
		#ID (k)	l_s(%)	T.1（初）	T.1k (s)	#ID (k)	l_s (%)	T.1（多个）	T.1k（复数）	#ID (k)	l_s(%)	T.1 (s)	T.1k (s)	#ID (k)	l_s (%)	T.1 (s)	T.1k (s)
人口普查	sex	10.4 (1.1)	10.3 (1.0)	**0.02 (ϵ)**	88.5 (13.1)	18.2 (1.0)	18.3 (0.1)	0.5 (ϵ)	52.1 (7.4)	21.6 (1.3)	19.5 (0.8)	0.05 (ϵ)	38.8 (6.5)	**79.0 (3.2)**	**74.0 (1.6)**	0.5 (ϵ)	**9.8 (0.6)**
	age	8.8 (7.2)	29.1 (2.2)	0.02 (ϵ)	113.5 (34.8)	21.6 (1.2)	55.8 (2.0)	0.5 (ϵ)	33.0 (7.3)	21.8 (11)	54.8 (1.8)	**0.01** (ϵ)	48.9 (19.7)	**112 (2.0)**	**93.4 (0.6)**	0.5 (ϵ)	**8.0 (1.0)**
	race	13.2 (9.7)	25.2 (1.5)	0.02 (ϵ)	89.9 (15.9)	24.1 (9.0)	43.3 (1.0)	0.7 (0.2)	40.6 (10.0)	25.9 (1.1)	43.5 (2.8)	**0.01** (ϵ)	37.5 (9.9)	**107 (1.7)**	**88.1 (0.8)**	0.5 (ϵ)	**8.8 (1.3)**
Compas	sex	12.6 (3.3)	13.8 (3.2)	**0.01** (ϵ)	118.2 (16.2)	17.3 (0.7)	17.4 (0.6)	0.02 (ϵ)	49.0 (14.0)	15.1 (1.1)	14.3 (0.9)	0.02 (ϵ)	56.3 (9.1)	**40.2 (1.5)**	**59.0 (1.4)**	0.2 (ϵ)	**27.6 (5.9)**
	age	6.8 (0.9)	10.3 (1.3)	0.02 (ϵ)	150.3 (23.7)	19.0 (0.9)	25.6 (1.2)	**0.01** (ϵ)	47.6 (5.8)	14.0 (0.8)	18.1 (0.9)	0.01 (ϵ)	70.1 (15.1)	**66.2 (3.0)**	**72.1 (1.6)**	0.01 (ϵ)	**19.7 (2.9)**
德语	sex	8.2 (0.8)	12.7 (1.2)	**0.01** (ϵ)	120.7 (20.1)	14.3 (0.7)	22.1 (0.8)	0.02 (ϵ)	68.7 (12.0)	12.9 (0.6)	18.3 (0.8)	0.09 (ϵ)	86.4 (14.5)	**62.6 (1.6)**	**80.2 (0.6)**	0.2 (0.2)	**15.0 (2.1)**
	age	8.4 (0.6)	39.8 (2.1)	0.02 (ϵ)	118.9 (22.9)	13.4 (1.1)	57.4 (3.0)	0.01 (ϵ)	60.6 (12.9)	12.9 (0.8)	52.5 (2.2)	**0.01** (ϵ)	56.5 (12.7)	**78.7 (0.8)**	**93.6 (0.2)**	0.05 (ϵ)	**11.1 (0.5)**
违约	sex	4.6 (0.6)	5.3 (0.6)	0.2 (0.2)	200.0 (45.1)	12.8 (0.7)	18.1 (1.1)	**0.01** (ϵ)	78.8 (7.3)	11.4 (0.9)	14.9 (1.3)	0.01 (ϵ)	80.0 (15.4)	**29.4 (1.5)**	**38.8 (1.9)**	2.1 (1.8)	**22.0 (10.5)**
	age	7.8 (2.5)	17.2 (3.2)	0.08 (ϵ)	130.4 (73.8)	13.3 (1.3)	43.4 (3.1)	**0.02** (ϵ)	78.9 (19.8)	9.0 (1.1)	29.9 (3.5)	0.06 (ϵ)	104.8 (14.7)	**50.4 (3.7)**	**68.4 (2.9)**	0.6 (ϵ)	**36.1 (17.1)**
心脏	sex	7.9 (3.2)	9.1 (3.4)	0.02 (ϵ)	171.8 (30.5)	10.6 (1.0)	13.6 (1.2)	**0.01** (ϵ)	94.9 (16.1)	10.2 (0.5)	10.9 (.5)	0.01 (ϵ)	84.4 (23.9)	**84.4 (1.3)**	**83.9 (1.2)**	0.04 (ϵ)	**7.7 (0.4)**
	age	16.7 (1.7)	48.1 (4.0)	0.02 (ϵ)	79.0 (13.2)	25.8 (1.7)	70.0 (2.5)	**0.01** (ϵ)	44.1 (18.6)	28.7 (1.2)	68.3 (1.4)	0.01 (ϵ)	37.9 (16.5)	**92.5 (1.6)**	**95.1 (0.4)**	0.04 (ϵ)	**7.5 (0.6)**
Bank	age	11.0 (1.8)	38.0 (4.1)	0.04 (ϵ)	115.0 (42.3)	7.9 (1.0)	34.2 (4.2)	**0.02** (ϵ)	136 (39.8)	9.2 (1.1)	36.6 (4.1)	0.03 (ϵ)	107.3 (40.2)	**47.4 (5.6)**	**87.3 (1.4)**	0.4 (0.2)	**25.0 (19.5)**
糖尿病	age	15.3 (1.0)	84.5 (1.7)	**0.01** (ϵ)	29.3 (4.7)	46.8 (0.9)	76.4 (0.8)	0.01 (ϵ)	17.8 (4.2)	47.9 (1.7)	76.2 (1.1)	0.01 (ϵ)	18.9 (4.7)	**171 (1.4)**	**94.2 (0.1)**	0.04 (ϵ)	**5.1 (0.2)**
学生	sex	2.4 (0.4)	6.1 (1.0)	0.03 (ϵ)	419.5 (58.9)	4.7 (0.4)	12.2 (1.0)	**0.01** (ϵ)	200.8 (25.5)	4.1 (0.6)	9.4 (1.3)	0.02 (ϵ)	246.5 (43.6)	**30.7 (1.1)**	**72.9 (1.3)**	1.2 (0.7)	**28.8 (4.0)**
	age	1.6 (0.3)	12.8 (2.3)	0.04 (ϵ)	564.8 (0.2k)	2.7 (0.4)	24.1 (2.5)	0.07 (ϵ)	330.9 (54.0)	2.7 (0.4)	20.8 (2.8)	**0.02** (ϵ)	409.7 (91.5)	**43.3 (1.0)**	**89.3 (0.5)**	0.5 (0.3)	**19.6 (1.4)**
MEPS15	sex	0.6 (0.1)	8.0 (1.4)	**0.01** (ϵ)	N/A	1.0 (0.2)	14.7 (2.2)	0.01 (ϵ)	337.0 (0.4k)	0.9 (0.2)	11.7 (2.3)	0.01 (ϵ)	**68.1 (0.2k)**	**6.2 (0.4)**	**67.9 (2.8)**	7.2 (6.4)	174.2 (33.5)
	age	0.4 (0.1)	16.1 (1.8)	0.03 (ϵ)	N/A	0.8 (0.2)	38.3 (6.8)	**0.02** (ϵ)	**92.0 (0.3k)**	0.7 (0.1)	29.1 (3.5)	0.05 (ϵ)	N/A	**6.7 (0.4)**	**82.1 (2.9)**	3.2 (3.4)	136.0 (12.7)
	race	0.7 (0.1)	9.4 (0.8)	**0.02** (ϵ)	N/A	1.4 (0.2)	21.1 (3.4)	0.02 (ϵ)	683.3 (0.1k)	1.0 (0.2)	14.2 (2.1)	0.02 (ϵ)	474.4 (0.4k)	**6.4 (0.4)**	**71.0 (2.2)**	13 (12)	**146.2 (25.8)**
MEPS16	sex	0.6 (0.1)	7.6 (1.2)	0.02 (ϵ)	N/A	1.2 (0.1)	17.6 (1.9)	**0.01** (ϵ)	605.2 (0.3k)	1.0 (0.1)	12.9 (1.8)	0.01 (ϵ)	256.0 (0.4k)	**5.8 (0.4)**	**74.1 (1.5)**	0.14 (ϵ)	**165.3 (19.3)**
	age	0.3 (ϵ)	13.9 (2.0)	0.02 (ϵ)	N/A	1.0 (0.2)	42.6 (7.9)	1.1 (1.1)	73.8 (0.2k)	1.0 (0.2)	38.1 (5.5)	0.03 (ϵ)	**65.5 (0.2k)**	**6.1 (0.3)**	**79.3 (1.6)**	1.4 (ϵ)	163.7 (12.9)
	race	1.1 (0.2)	14.5 (2.1)	0.01 (ϵ)	654.4 (0.3k)	1.7 (0.2)	26.3 (3.8)	**0.01** (ϵ)	509.3 (0.1k)	2.1 (0.3)	27.1 (2.7)	0.01 (ϵ)	480.4 (49.1)	**6.3 (0.4)**	**76.9 (2.7)**	0.2 (ϵ)	**163.6 (20.2)**

表 IV：对因果影响公平性的层和神经元进行定位。Neuron_i^+ 表示激活后对公平性产生正面影响的排名第 i 的神经元索引；Neuron_i^- 表示激活后对公平性产生负面影响的排名第 i 的神经元索引；ACD_i^+ 表示对公平性产生正面影响的排名第 i 的神经元的（归一化）平均因果差异；ACD_i^- 表示对公平性产生负面影响的排名第 i 的神经元的（归一化）平均因果差异。

数据集	层索引	层影响	神经元$_1^+$	ACD$_1^+$	神经元$_2^+$	ACD$_2^+$	神经元$_3^+$	ACD$_3^+$	神经元$_1^-$	ACD$_1^-$	神经元$_2^-$	ACD$_2^-$	神经元$_3^-$	ACD$_3^-$
人口普查	2	9.01	N_{19}	0.159	N_3	0.150	N_{12}	0.122	N_{15}	0.183	N_{24}	0.163	N_{14}	0.102
Compas	2	20.70	N_{25}	0.516	N_{28}	0.419	N_7	0.346	N/A	N/A	N/A	N/A	N/A	N/A
德语	5	1.79	N_3	0.050	N_1	0.013	N/A	N/A	N_0	0.042	N/A	N/A	N/A	N/A
违约	2	27.58	N_5	0.039	N_{27}	0.022	N_{19}	0.014	N_2	0.031	N_{13}	0.027	N_{10}	0.019
心脏	4	1.47	N_5	0.120	N_4	0.048	N_0	0.044	N/A	N/A	N/A	N/A	N/A	N/A
Bank	3	6.62	N_0	0.495	N_7	0.178	N_1	0.091	N_6	0.057	N_{11}	0.014	N/A	N/A
糖尿病	2	1.67	N_{19}	0.041	N_6	0.035	N_5	0.031	N_0	0.042	N_{29}	0.001	N/A	N/A
学生	2	4.01	N_{22}	0.550	N_{24}	0.442	N_8	0.229	N_4	0.084	N_{18}	0.055	N_{28}	0.026
MEPS15	2	35.44	N_{24}	0.230	N_6	0.167	N_{14}	0.160	N/A	N/A	N/A	N/A	N/A	N/A
MEPS16	2	47.46	N_8	0.147	N_{11}	0.138	N_{24}	0.114	N_{30}	0.006	N_{30}	0.003	N_{22}	0.001

表V： A 表示准确率， K 表示测试用例的平均簇数； $A^{=0}$ 表示停用公平性影响最大的负向神经元后的准确率； $K^{=0}$ 表示停用后的平均簇数； $A^{>0}$ 表示激活公平性影响最大的正向神经元后的准确率； $K^{>0}$ 表示激活后的平均簇数； T_I 表示定位和缓解所需的计算时间（秒）。

数据集	A	K	A$^{=0}$	K$^{=0}$	A$^{>0}$	K$^{>0}$	T$_I$
人口普查	0.882	22.68	0.85	18.22	0.85	18.68	1,538
Compas	0.976	5.19	N/A	N/A	0.971	2.66	1，226
德语	1.0	5.05	0.992	4.62	0.956	3.53	559
违约	0.827	8.08	0.822	6.70	0.826	7.70	1,230
心脏	0.96	6.91	N/A	N/A	0.916	6.59	553
Bank	0.923	5.95	0.901	4.32	0.893	3.16	858
糖尿病	0.993	3.90	0.956	2.91	0.965	3.61	1,238
学生	1.0	6.42	0.97	5.79	0.98	2.29	1，206
MEPS15	0.898	9.35	N/A	N/A	0.866	6.93	1,348
MEPS16	0.913	7.42	0.906	6.98	0.903	6.27	1,368

目前尚不确定这些数据集和深度神经网络模型是否足以代表公平性测试所需的情况。

VIII. 相关工作

机器学习系统的公平性测试。THEMIS [10]提出了一种因果歧视概念，通过比较两个子组的公平性指标差异来反事实查询；即，它们对受保护属性设为 A 的个体进行采样，并将结果与受保护属性设为B 的反事实场景进行比较。符号生成（SC） [28], [22] 提出了一种黑盒测试方法，该方法使用决策树近似 ML 模型，并利用树结构上的符号执行来发现个体歧视（ID）。AEQUITAS [19] 采用两步方法：首先从输入数据集中均匀随机地抽取实例，以找到一个歧视性实例；然后在局部扰动这些实例，以进一步生成有偏差的测试用例。EXPGA [16] 提出了一种遗传算法（GA），用于在自然语言处理中生成 ID实例。该技术使用先验知识图谱来指导 NLP 任务中受保护属性的扰动。虽然这些技术是黑盒的，但它们可能会因搜索过程中缺乏局部指导而受到影响。ADF [17]利用损失函数的梯度作为生成 ID 实例的指导。全局阶段探索输入空间，以找到多样化的个体歧视集合，而局部阶段则利用每个实例，在其邻域中生成许多个体歧视（ID）实例。EIDIG [29]遵循与ADF类似的思路，但使用不同的梯度计算方式。首先，它使用输出的梯度（而非损失函数的梯度），以降低每次迭代的计算成本。其次，它使用梯度的动量来

全局阶段以避免局部最优解。NEURONFAIR [18]扩展 ADF 和EIDIG，以支持非结构化数据（例如图像、文本、语音等），其中受保护属性可能没有明确定义。此外，NEURONFAIR由DNN 的内部神经元状态（例如激活和停用的模式）及其激活差异引导。除了这些技术的能力之外，DICE 还量化歧视程度，使软件开发人员能够确定测试用例的优先级，并同时搜索多个受保护属性。

超出本文的讨论范围，已有大量研究 [42], [43], [44], [23], [45], [46] 考虑了群体公平性测试。FAIRWAY [43] 在找到合适的机器学习算法配置后缓解偏差。为此，他们采用了多目标优化方法（FLASH） [47]。PARFAIT-ML [45]通过灰盒进化算法搜索经典机器学习算法的超参数空间，以刻画超参数配置所导致偏差的程度。

深度神经网络的调试。CRADLE [33]追踪了DNN模型在两种不同深度学习框架上的执行图，并利用结果差异定位可能导致错误的后端函数。然而，由于CRADLE未使用因果分析，因此假阳性率较高。AUDEE [48] 采用了类似的方法，但利用了因果测试方法。具体而言，它设计了干预DNN模型的策略，并追踪干预对观测到的不一致性的影响。我们改编了CRADLE和AUDEE的层定位方法；但我们的因果定位是基于针对元属性（公平性）的do逻辑开发的。AUDEE通过对神经元值进行简单扰动来检验功能正确性（即任何不一致都表明存在错误），但未考虑准确率或神经元对错误的影响程度。

进程内缓解。一系列研究探讨了用于缓解机器学习预测中偏差的进程内算法 [49], [50], [51]。对抗式去偏 [49]和偏见消除器 [50]通过向模型参数或损失函数添加约束来提升公平性。指数梯度 [51] 采用元学习算法，推断出一组能够最大化准确率与公平性的分类器。与这些方法不同，我们开发了一种专门用于处理个体公平性神经网络的缓解方法。这一设定使我们能够利用神经网络基于层的结构进行因果推理和缓解。我们认为，我们的方法可以与进程内缓解技术结合，从而最大化基于深度神经网络的决策支持系统中的公平性。

形式化方法。我们相信，本文能够与形式验证及其应用领域的丰富文献建立联系。下面给出两个例子。Fairsquare [52] 使用一种称为加权体积计算算法的新型验证技术，对概率程序中的公平决策过程进行认证。SFTREE [53]将公平决策树推断问题表述为混合整数线性规划问题，并迭代地应用约束求解器来寻找解决方案。

收入、财富和税收中的公平性。我们开发了一种专门用于处理回归问题的公平性测试与调试方法。因此，我们的方法有助于研究和解决不同种族和性别群体 [54] 之间收入和财富分配中的偏差。此外，我们的方法也有助于研究税收中的公平性（例如，纵向和横向公平性 [55], [56]）。我们将这些方向的进一步研究留待未来工作。

IX. 结论

基于深度神经网络的软件解决方案正越来越多地应用于社会关键领域，其设计中的错误可能导致歧视性行为。在本文中，我们介绍了DICE：一种信息论模型，用于刻画基于深度神经网络的决策所使用的受保护信息量。我们的实验表明，基于定量空间的搜索和调试算法能够有效地发现并定位公平性缺陷。

致谢。作者感谢匿名的 ICSE 审稿人拨冗审阅，并提出宝贵意见以改进本文。本研究部分得到了美国国家科学基金会（NSF）项目（资助编号：DGE-2043250）以及 UTEP 工程学院创业启动经费的支持。

参考文献

[1] I. J. Goodfellow、Y. Bengio 和 A. Courville，《深度学习》。美国马萨诸塞州剑桥：MIT 出版社，2016 年，http://www.deeplearningbook.org。[2] S. Ranchord´as 和 L. Scarcella，《面向弱势公民的自动化政府：权利中介》，*SSRN* 电子期刊，2021 年。[3] N. Tishby、F. C. Pereira 和 W. Bialek，《信息瓶颈方法》，*arXiv* 预印本 *physics/0004057*，2000 年。[4] S. M. Julia Angwin、Jeff Larson 和 L. Kirchne，《机器偏差》，https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing，2021 年，在线。[5] D. A. Elyounes，《“ 计算机说不！ ”：自动化对公职人员自由裁量权的影响》，范德比尔特娱乐与科技法律杂志，第 23 卷，第 451 页，2020 年。[6] N. Escher 和 N. Banovic，《揭示贫困管理技术中的错误：审计政府福利筛查工具的方法》，*ACM* 人机交互会议论文集，第 4 卷，CSCW 期，第 064:1– 064 20, 2020 [ ] 10 1145 3392874 : 。[7] 在线。可查阅：https://doi.org/ 。/ J. Slemrod，《税收制度中的群体公平与隐性歧视》，《国家税务杂志》，第 75 卷，第 1 期，第 201–224 页，2022 年。[8] D. A. Brown，《美国国税局正在针对最贫困的美国人》，8 月 2021， [在线；发布于 27-July-2021]。[在线]。可查阅：https://www.theatlantic.com/ideas/archive/2021/07/how-race-plays-tax-policing/619570/ [9] C. Thomas 和 A. Pont´on-N´u˜nez，《自动化司法裁量：庭前裁决中的算法风险评估如何基于种族侵犯平等保护权》，《明尼苏达州法律与不平等杂志》，第 40 卷，第 2 期，第 371 页，2022 年。[10] R. Angell、B. Johnson、Y. Brun 和 A. Meliou，《Themis：自动测试软件中的歧视》，载于*2018* 年第 *26* 届 *ACM* 欧洲软件工程会议与软件工程基础研讨会联合会议论文集，2018 年，第 871–875 页。[11] S. Galhotra、Y. Brun 和 A. Meliou，《公平性测试：测试软件中的歧视》，载于*FSE*，2017 年，第 498–510 页。[12] A. Aggarwal、P. Lohia、S. Nagar、K. Dey 和 D. Saha，《机器学习模型的黑盒公平性测试》，载于 *2019* 年第 *27* 届 *ACM* 欧洲软件工程会议与软件工程基础研讨会联合会议论文集，2019 年，第 625–635 页。

[13] A. Sharma 和 H. Wehrheim，“机器学习模型的自动化公平性测试”，载于 IFIP 国际软件与系统测试会议。Springer，2020 年，第 255–271 页。[14] M. Hardt、E. Price 和 N. Srebro，“监督学习中的机会平等”，载于 NIPS，2016 年。[15] C. Dwork、M. Hardt、T. Pitassi、O. Reingold 和 R. Zemel，“通过认知实现公平性”，载于第 3 届理论计算机科学创新会议论文集，2012 年，第 214–226 页。[16] M. Fan、W. Wei、W. Jin、Z. Yang 和 T. Liu，“通过遗传算法进行解释引导的公平性测试”，载于第 44 届国际软件工程会议论文集，系列：ICSE '22。美国纽约州纽约市：计算机协会，2022 年，第 871 882 [ ] 10 1145 3510003 3510137 – 页。[17] 在线。获取地址：https://doi.org/ 。 / 。P. Zhang、J. Wang、J. Sun、G. Dong、X. Wang、X. Wang、J. S. Dong 和 T. Dai，“通过对抗性采样进行白盒公平性测试”，载于 ACM/IEEE 第 42 届国际软件工程会议论文集，2020 年，第 949–960 页。[18] H. Zheng、Z. Chen、T. Du、X. Zhang、Y. Cheng、S. Ti、J. Wang、Y. Yu 和 J. Chen，“NEURONFAIR：通过有偏神经元识别进行可解释的白盒公平性测试”，载于 2022 年 IEEE/ACM 第 44 届国际软件工程会议（ICSE），2022 年，第 1519–1531 页。[19] S. Udeshi、P. Arora 和 S. Chattopadhyay，“自动化定向公平性测试”，载于第 33 届 ACM/IEEE 国际自动化软件工程会议论文集，2018 年，第 98–108 页。[20] M. Glymour、J. Pearl 和 N. P. Jewell，《统计学中的因果推断：入门》，John Wiley & Sons，2016 年。[21] Q. Guo、X. Xie、Y. Li、X. Zhang、Y. Liu、X. Li 和 C. Shen，“Audee：深度学习框架的自动化测试”，载于 2020 年第 35 届 IEEE/ACM 国际自动化软件工程会议（ASE）。IEEE，2020 年，第 486–498 页。[22] A. Agarwal、P. Lohia、S. Nagar、K. Dey 和 D. Saha，“自动生成测试以检测人工智能模型中的个体歧视”，arXiv 预印本 arXiv:1809.03260，2018 年。[23] J. Chakraborty、S. Majumder 和 T. Menzies，“机器学习软件中的偏差：为什么？如何？该怎么办？”，载于第 29 届 ACM 欧洲软件工程会议与软件工程基础研讨会联合会议论文集，系列：ESEC/FSE 2021。美国纽约州纽约市：计算机协会，2021 年，第 429 440 [ ] 10 1145 3468264 3468537 – 页。[24] 在线。获取地址：https://doi.org/ 。 / 。A. Renyi 等，“关于熵与信息的度量”，载于第四届伯克利数学统计与概率研讨会论文集，第 1 卷，第 547–561 页。美国加利福尼亚州伯克利，1961 年。[25] G. Smith，“定量信息流的基础”，载于《软件科学与计算结构基础》，L. de Alfaro 编。德国柏林、海德堡：Springer Berlin Heidelberg，2009 年，第 288–302 页。[26] M. Backes、B. K¨opf 和 A. Rybalchenko，“信息泄漏的自动发现与量化”，载于 2009 年第 30 届 IEEE 安全与隐私研讨会。IEEE，2009 年，第 141–153 页。[27] Y. Noller 和 S. Tizpaz-Niari，“Qfuzz：针对侧信道的定量模糊测试 ”，系列：ISSTA 2021。美国纽约州纽约市：计算机协会，2021 年，第 257–269 页。 [在线]。获取地址：https://doi.org/ 10 1145 3460319 3464817 。 / 。A. Aggarwal、P. Lohia、S. Nagar、K. Dey 和 D. Saha，“机器学习模型的黑盒公平性测试”，载于 2019 年第 27 届 ACM 欧洲软件工程会议与软件工程基础研讨会联合会议论文集，系列：ESEC/FSE 2019，2019 年，第 625–635 页。[29] [ ] 10 1145 3338906 3338937 在线。获取地址：https://doi.org/ 。 / 。L. Zhang、Y. Zhang 和 M. Zhang，“通过梯度搜索进行高效白盒公平性测试”，载于第 30 届 ACM SIGSOFT 国际软件测试与分析研讨会论文集，系列：ISSTA 2021，2021 年，第 103–114 页。 [在线]。获取地址：https://doi.org/ 10 1145 3460319 3464820 。 / 。J. Pearl，《因果关系》，剑桥大学出版社，2009 年。[31] D. Dua 和 C. Graff，“加州大学欧文分校机器学习库”，2017 年。 [在线]。获取地址：https://archive.ics.uci.edu/ml/datasets/census+income [32] D. C. Liu 和 J. Nocedal，“大型优化问题的有限内存 BFGS 方法”，《数学规划》，第 45 卷，第 1 期，第 503–528 页，1989 年。[33] H. V. Pham、T. Lutellier、W. Qi 和 L. Tan，“Cradle：通过跨后端验证检测并定位深度学习库中的错误”，载于 2019 年 IEEE/ACM 第 41 届国际软件工程会议（ICSE）。IEEE，2019 年，第 1027–1038 页。[34] ProPublica，《Compas 软件分析》，https://github.com/propublica/compas-analysis，2021，在线。[35]《UCI：Statlog（德国信用数据）数据集》，2000 年。 [在线]。可访问：https://archive.ics.uci.edu/ml/datasets/statlog+（德国+信用+数据）[36]《UCI：信用卡客户违约数据集》，2009 年。 [在线]。可访问：https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients [37]《UCI：心脏病数据集》，2001 年。 [在线]。可访问：https: //archive.ics.uci.edu/ml/datasets/Heart+Disease[38]《银行营销 UCI》，2017 年。 [在线]。可访问：https://archive.ics.uci. edu/ml/datasets/bank+marketing [39]《UCI：糖尿病患者记录》，1994 年。 [在线]。可访问：https: //archive.ics.uci.edu/ml/datasets/diabetes[40]《学生表现数据集》，2014 年。 [在线]。可访问：https: //archive.ics.uci.edu/ml/datasets/Student+Performance[41]《医疗支出专题调查》，2014 年。 [在线]。可访问：https://meps.ahrq.gov/mepsweb/ [42] R. K. Bellamy、K. Dey、M. Hind、S. C. Hoffman、S. Houde、K. Kannan、P. Lohia、J. Martino、S. Mehta、A. Mojsilovi´c 等，《AI 公平性 360：用于检测和缓解算法偏见的可扩展工具包》，《IBM 研究与开发杂志》，第 63 卷，第 4/5 期，第 4–1 页，2019 年。[43] J. Chakraborty、S. Majumder、Z. Yu 和 T. Menzies，《Fairway：构建公平的机器学习软件的方法》，载于第 28 届 ACM 欧洲软件工程会议与软件工程基础研讨会联合会议论文集，2020 年，第 654–665 页。[44] J. M. Zhang 和 M. Harman，《软件公平性中的“无知与偏见”》，载于第 43 届 IEEE/ACM 国际软件工程会议，ICSE 2021，西班牙马德里，2021 年 5 月 22–30 日。IEEE，2021 年，第 1436–1447 页。 [在线]。可访问：https: //doi.org/ 10 1109 43902 2021 00129 . /ICSE . . S. Tizpaz-Niari、A. Kumar、G. Tan 和 A. Trivedi，《机器学习库的公平性感知配置》，载于第 44 届 IEEE/ACM 国际软件工程会议，ICSE 2022，美国宾夕法尼亚州匹兹堡，2022 年 5 月 25–27 日。IEEE，2022 年，第 909–920 页。[45] [ ] 10 1145 3510003 3510202 在线。可访问：https://doi.org/ . / . Z. Chen、J. M. Zhang、F. Sarro 和 M. Harman，《Maat：解决机器学习软件公平性与性能缺陷的新型集成方法》，丛书：ESEC/FSE 2022，2022 年，第 1122–1134 页。[46] [ ] 10 1145 3540250 3549093 在线。可访问：https://doi.org/ . / . V. Nair、Z. Yu、T. Menzies、N. Siegmund 和 S. Apel，《使用 Flash 查找更快的配置》，《IEEE 软件工程汇刊》，第 46 卷，第 7 期，第 794–811 页，2020 年。[47] Q. Guo、X. Xie、Y. Li、X. Zhang、Y. Liu、X. Li 和 C. Shen，《Audee：深度学习框架的自动化测试》，载于 2020 年第 35 届 IEEE/ACM 自动化软件工程国际会议（ASE），2020 年，第 486–498 页。[48] B. H. Zhang、B. Lemoine 和 M. Mitchell，《使用对抗学习缓解不必要的偏差》，载于 2018 年 AAAI/ACM 人工智能、伦理与社会会议论文集，2018 年，第 335–340 页。[49] F. Kamiran、A. Karim 和 X. Zhang，《歧视感知分类的决策理论》，载于 2012 年 IEEE 第 12 届国际数据挖掘会议，2012 年，第 924–929 页。[50] A. Agarwal、A. Beygelzimer、M. Dud´ık、J. Langford 和 H. Wallach，《公平分类的规约方法》，载于国际机器学习会议。机器学习研究论文集，2018 年，第 60–69 页。[51] A. Albarghouthi、L. D'Antoni、S. Drews 和 A. V. Nori，《Fairsquare：程序公平性的概率验证》，第 OOPSLA 号，2017 年。[52] [ ] 10 1145 3133904 在线。可访问：https://doi.org/ . / J. Wang、Y. Li 和 C. Wang，《通过迭代约束求解综合公平决策树》，载于《计算机辅助验证：第 34 届国际会议》，CAV 2022，以色列海法，2022 年 8 月 7–10 日，论文集，第二部分。Springer，2022 年，第 364–385 页。[53] [54] T. Piketty 和 E. Saez，《美国的收入不平等，1913–1998，41，2003 118，1，1》。《经济学季刊》，第 卷，第 期，第 – 页。[55]E. Black、H. Elzayn、A. Chouldechova、J. Goldin 和 D. Ho，《算法公平性与垂直公平：基于美国国税局税务审计模型的收入公平》，载于 2022 年 ACM 公平性、问责制与透明度会议，丛书：FAccT '22。美国纽约州纽约市：计算机协会，2022 年，第 1479–1503 页。 [在线]。可访问：https://doi.org/ 10 1145 3531146 3533204 . / . S. Tizpaz-Niari、V. Monjezi、M. Wagner、S. Darian、K. Reed 和 A. Trivedi，《税务申报软件的蜕变测试与调试》，2023 年。 [在线]。可访问：https://arxiv.org/abs/2205.04998 [56]