文章编号:1003 - 2053(2026)05 - 0982 - 10

# 算法公平定义的困境与出路
## ——从统计公平到因果公平

胡嘉伟
(南京大学哲学学院,江苏南京 210023)

**摘 要**:人工智能算法在各种决策场景的重要性日益凸显,但同时面临决策公平性的挑战。当前,算法公平的主流定义进路是统计公平进路,该进路通过设定具体的统计指标来衡量算法的公平性。然而,统计公平进路因其理论局限面临一些难以解决的困境。因此,算法公平性研究逐渐趋向基于"因果范式"的因果公平进路。因果公平进路不仅能够较好地回应统计公平进路所面临的困境,而且能适配更好的算法公平性方法论。此外,从统计公平到因果公平的研究趋向,体现了人工智能可解释性、透明性、可信赖性的核心诉求。

**关键词**:算法公平;统计公平;因果公平;结构因果模型;方法论
**中图分类号**:NO31 **文献标识码**:A

随着人工智能技术的发展,越来越多场景开始应用人工智能作为重要的决策工具,比如金融借贷、医疗诊断、司法量刑以及职场招聘。然而,人工智能辅助决策虽然能够极大提高人们的工作效率,但也带来许多风险与挑战。其中,人工智能算法决策的"公平(fairness)"风险备受关注。2016 年,用于预测罪犯未来是否再犯的机器学习系统 COMPAS 被曝出对黑人罪犯群体有着严重偏见和歧视[1]。2018 年,亚马逊的人工智能招聘系统也被发现对女性应聘者有偏见[2]。更有甚者,在医学图像分析领域,部分深度学习算法的诊断性能存在显著的群体差异,具体表现为对女性、黑人以及低社会经济群体的诊断准确性明显低于男性、白人及高社会经济群体[3]。除此之外,保险定价、电商杀熟等场景也存在公平风险。

这些案例不仅引起了社会对算法公平的广泛关注,同时也点燃了人们对人工智能的信任危机。在此背景下,不少国家在人工智能相关的政策文件中将公平作为核心要求之一,比如我国提出的《全球人工智能治理倡议》、欧盟颁布的《人工智能法案》以及美国 2025 年发布的《关于人工智能未来研究报告》等。由此可见,人工智能算法的公平诉求已成为全球人工智能治理的广泛共识。

当前,国内学界对算法公平或算法歧视主要从技术细节、伦理讨论和治理对策几个大方面展开研究,却鲜有对最根本的"算法公平定义"做技术哲学反思。算法公平定义是用于衡量算法是否公平的关键标准,同时它也可作为公平约束嵌入算法内部,使算法的输出结果达到某种程度的公平性。可以说,几乎所有的算法公平研究都要基于公平定义展开。然而,由于"公平"这一概念极具深刻内涵,且不同文化历史情境下的公平标准可能各不相同,因此衍生出了多种多样的定义。总体来看,我们可以根据定义范式的不同,将算法公平定义研究分为两类——统计公平与因果公平。统计公平的研究进路采取以数据为中心的"统计范式"开展,易于操作,如今占据主流的应用场景;而因果公平进路则采取以数据解释为中心的"因果范式"开展,需要构建数据背后的因果假设,如今也呈现出愈发重要的态势。

本文旨在探究当前算法公平定义研究的现状、困境与趋向。第一部分简要考察算法公平定义的主流进路,即统计公平进路;第二部分分析统计公平进路面临的三大困境,并指出这些困境因其理论范式的局限性而无法在该范式下解决;第三部分讨论因果公平进路如何能应对统计公平进路的困境;论文最后探讨了算法公平研究的方法论问题,相比统计

**收稿日期**:2025 - 04 - 29;**修回日期**:2025 - 09 - 23
**基金项目**:江苏省社会科学基金重大项目(24ZD006)
**作者简介**:胡嘉伟(1997 - ),男,博士研究生,E - mail:hjwnju@ foxmail. com。

公平进路,因果公平进路对应了实现算法公平的更好方法论,同时,因果公平进路还能助力解决人工智能的可解释性问题,提高算法的可信赖性。

## 1　算法公平定义的统计进路

算法公平有一个较为通行的说明性定义:“在决策情境中,对个体或群体不存在因其固有或后天的特征所引起的偏见或偏袒”[4]。其中,“固有或后天的特征”也被称作“敏感属性”或“受保护属性”,常见的受保护属性有种族、性别、年龄等。在某些特定场景中,如防止电商价格歧视,消费记录可能也会作为受保护属性[5]。一般来说,如果某个算法因群体或个体具有不同受保护属性而采取不同的决策行为,那么该算法就是不公平的。

统计公平进路作为算法公平定义的主流进路,以各项统计数据指标来衡量公平性,其核心思想可简单概括为:如果算法决策结果与受保护属性在某一指标的条件下保持统计独立,那么该算法满足该指标意义上的公平性。大体上,统计公平进路对公平的定义可以分为三类,“统计均等”“精度统计均等”以及“校准公平”。

### 1.1　统计均等

统计均等(statistical parity)或人群均等(demographic parity)[6]是统计公平进路中最基本的一种定义,该定义要求算法在根据受保护属性划分的不同群体中,产生相同概率的输出结果,即算法结果与受保护属性的值统计独立。

以某累犯预测算法为例。假设“种族(变量 $A$)”是该算法唯一的受保护属性,当 $A=1$ 时表示白人罪犯(非弱势群体),$A=0$ 时表示黑人罪犯(弱势群体)。标签 $\hat{Y}$ 代表算法的预测输出结果,$\hat{Y}=1$ 为高风险再犯,$\hat{Y}=0$ 为低风险再犯①。统计均等要求可用下述条件概率等式表征:

$$P(\hat{Y}=\hat{y}\mid A=0)=P(\hat{Y}=\hat{y}\mid A=1)。$$

也就是说,如果按照统计均等来衡量该算法的公平性,黑人罪犯和白人罪犯群体被算法预测为高风险或低风险再犯的概率相等时,该算法才是公平的。

值得注意的是,所有算法公平定义都是一种理想化的数学标准设定,现实生活中难以得到以上两个完全相等的概率值。因此,我们可以通过不同概率值的差值(的绝对值)或比值是否在合理的阈值范围来衡量该算法的公平性。根据统计均等定义,两个不同群体对应的概率值之比通常称为“差异影响(disparate impact)”。比如,美国某些法律原则,要求招聘场景中受保护的弱势群体获得工作的概率与非弱势群体获得工作概率之比不应低于 80%,这也被称为“80% 原则(或 4/5 原则)”[7]。统计均等定义适用于对算法输出结果有明确不同群体比例要求的场景。

### 1.2　精度统计均等

在统计均等的基础上,引入“实际结果标签 $Y$”为条件,就构成了精度统计均等这一类定义。实际结果标签 $Y$ 表示个体在现实世界中实际发生的结果,比如“(两年内)是否再次犯罪”“是否按时还贷款”等等,这些实际结果通常源于历史数据或是跟踪观察所得。精度统计均等把算法预测结果($\hat{Y}$)与实际结果($Y$)的(不)相符程度在不同群体中做比较,以此来更精细地衡量公平性。

我们一般把预测结果 $\hat{Y}=1$ 看作“阳性”,$\hat{Y}=0$ 看作“阴性”。当预测结果与实际结果 $Y$ 的值相符时为“真”,不相符时则为“假”,由此可以排列出下表所示的四种情况。

**表 1　预测结果与实际结果的组合情况**

**Table 1　Combination of predicted and actual outcomes**

| 预测($\hat{Y}$)\实际($Y$) | $Y=1$ | $Y=0$ |
|---|---|---|
| $\hat{Y}=1$ | 真阳(true positive/TP) | 假阳(false positive/FP) |
| $\hat{Y}=0$ | 假阴(false negative/FN) | 真阴(true negative/TN) |

相应地,以上 4 种情况也对应了 4 种比率,这 4 种比率分别代表了 4 种公平指标,我们仍以上述累犯预测算法为例,总结为下表所示。

根据以上 4 种公平指标,可以得到几个最为常用的精度统计均等定义,即几率均等(equalized odds)、机会均等(equal opportunity)[8]和“待遇均等(treatment equality)”[9]。几率均等要求算法对于不同群体预测结果都满足真阳率和假阳率均等;机会均等是弱化的几率均等,只要求真阳率均等;待遇均等则要求假阳率和假阴率均等。

有两点值得注意:第一,精度统计均等与统计均等在数学形式上的主要区别在于,条件概率式中增

① 为方便说明,本文的受保护属性、结果标签均为二值变量,即取值 0 或 1。

加了以“实际结果 $Y$”为条件,其意义是呈现算法“预测结果 $\hat{Y}$”在“实际结果 $Y$”中的人群占比;第二,真阳率与假阴率、真阴率与假阳率这两对指标之和为 1,这一方面意味着它们之间是此消彼长的,另一方面意味着真阳率均等满足的同时假阴率均等随即满足(真阴率与假阳率同理)。

**表 2 精度统计均等的指标、意义及示例**

**Table 2 Metrics, significance, and examples of statistical parity in accuracy**

| 比率 | 公平指标 | 公平意义 | 示例 |
|---|---|---|---|
| 真阳率 TPR | $TPR = TP/(TP + FN)$<br>真阳率均等:<br>$P(\hat{Y}=1 \mid A=0, Y=1) = P(\hat{Y}=1 \mid A=1, Y=1)$ | 算法对正例群体($Y=1$)的对判率 | 实际再犯的罪犯被正确预测为高风险的概率 |
| 假阳率 FPR | $FPR = FP/(FP + TN)$<br>假阳率均等:<br>$P(\hat{Y}=1 \mid A=0, Y=0) = P(\hat{Y}=1 \mid A=1, Y=0)$ | 算法对负例群体($Y=0$)的错判率 | 实际不再犯的罪犯被错误预测为高风险的概率 |
| 真阴率 TNR | $TNR = TN/(TN + FP)$<br>真阴率均等:<br>$P(\hat{Y}=0 \mid A=0, Y=0) = P(\hat{Y}=0 \mid A=1, Y=0)$ | 算法对负例群体($Y=0$)的对判率 | 实际不再犯的罪犯被正确预测为低风险的概率 |
| 假阴率 FNR | $FNR = FN/(FN + TP)$<br>假阴率均等:<br>$P(\hat{Y}=0 \mid A=0, Y=1) = P(\hat{Y}=0 \mid A=1, Y=1)$ | 算法对正例群体($Y=1$)的错判率 | 实际再犯的罪犯被错误预测为低风险的概率 |

### 1.3 校准公平

校准公平(calibration fairness)或测试公平(test fairness)[10]也是有实际结果标签 $Y$ 参与的一类公平定义。它们与精度统计均等的区别在于,实际结果标签 $Y$ 不是作为条件出现的,而是反过来以预测结果 $\hat{Y}$ 为条件。最常见的校准公平定义称为预测均等(predictive parity),预测均等可以看作几率均等的逆向条件[11],可用以下等式表征:

$P(Y=1 \mid A=0, \hat{Y}=\hat{y}) = P(Y=1 \mid A=1, \hat{Y}=\hat{y})$。

该定义以算法预测结果 $\hat{Y}$ 为条件,通过判断不同群体的预测准确率来衡量公平性。比如,如果算法对于黑人罪犯和白人罪犯的再犯行为的预测准确率是相等的,那我们就称该算法满足预测均等。换句话说,预测均等追求的是算法的预测结果尽量能贴近不同人群的实际结果。比如黑人罪犯和白人罪犯在社会上的实际再犯率分别是 $P(Y=1 \mid A=0)$ 和 $P(Y=1 \mid A=1)$,如果算法对黑人罪犯和白人罪犯的预测是符合实际再犯率的,那么该算法满足预测均等。由于该定义是将算法预测结果相对于实际结果进行“校正”,因此属于某种校正公平。

以上定义都属于统计公平定义,它们具有三个共同特点。第一,它们都将“公平性”转换为可计算的数学条件,如概率、比率等统计量之间的差值关系。如此在数学上容易刻画与计算,且便于应用到目前主流的机器学习算法中,技术实现难度较小;第二,它们都关注不同群体之间的差异,在较为宏观的层面衡量公平性。如此定义虽然有利于群体公平性,但对群体内的个体不公无能为力;第三,它们都基于算法的决策结果作为公平性约束,换句话说,它们只看重结果是否符合某种统计指标约束,而不在意算法内部如何产生这样的结果。

统计公平之所以成为主流进路,与上述第一个特点的关系最为密切。当前主要的机器学习算法,如深度学习,根本上是基于数据的统计范式开展的。因此,统计公平进路能够很自然地与机器学习算法相结合。统计公平虽有一些优势,但上述另两个特点也揭示了该进路的局限性。下文将表明,统计公平进路还面临着三个无法解决的困境,导致这些困境的根源不是还未找到更好、更合适的统计定义,而是统计范式自身的根本性缺陷。

## 2 统计公平进路的三大困境

统计公平进路至少面临三大困境,它们分别是“定义非兼容性困境”“公平悖谬困境”和“实际标签偏见困境”。这三大困境受制于统计范式本身的缺陷,无法在统计公平进路中根本解决,破局的关键是

转变定义公平的范式。

### 2.1　定义非兼容性困境

统计公平进路面临定义非兼容性的困境,“定义非兼容性”是指部分重要的统计公平定义之间存在数学意义上的冲突,无法同时满足。比如,克莱因伯格(Jon Kleinberg)等人证明了对于预测均等、假阳率均等以及假阴率均等这三种定义来说,除非在高度受限的特殊情况,否则它们三者无法同时满足,即使是近似满足也不可能[12]。丘尔德乔娃(Alexandra Chouldechova)则进一步证明,当不同群体之间基本率(base rate)不同时②,预测均等甚至无法与假阳率均等或者假阴率均等的二者之一兼容。[10]许多学者将这一结论称作“公平的不可能性(the impossibility of fairness)”[13]。

这一现象之所以导致困境,是因为在某些场景中,不同统计公平定义的支持者们会产生强烈分歧,但又无法同时满足。在 2016 年累犯预测的 COMPAS 案例中,该算法被媒体曝出黑人罪犯的假阳率,即实际未再犯的低风险黑人被算法“冤枉”的概率,要比白人高出两倍之多(45%,23%);与此同时,白人罪犯的假阴率,即实际再犯的高风险白人被算法“放过”的概率,比黑人高出很多(48%,28%)[14]。这两项数据似乎说明了该算法对黑人罪犯很严苛,但对白人罪犯却很宽松。这样看来,COMPAS 远远达不到精度统计公平中的待遇均等(假阳率与假阴率均等)。当大多数人都依照这两项数据指控 COMPAS 的算法不公时,一些研究者却提出了反对意见。迪特里希(William Dieterich)等人认为应该从预测均等的角度来衡量 COMPAS 的公平性[15]。根据预测均等定义,被算法预测为高风险的黑人罪犯实际再犯的概率基本与白人罪犯相当,甚至略高于白人罪犯(63%,59%)。也就是说,从算法预测结果的准确率来看,COMPAS 并没有歧视黑人罪犯,而是做到了基本贴近这两个群体的实际表现。

丘尔德乔娃所证明的“公平的不可能性”结论告诉我们,只要黑人和白人群体实际的再犯率(基本率)不同,我们就不可能使算法能同时满足预测均等和假阳率均等(或假阴率均等)。但是不同的定义支持者们各不相让,我们又该如何评定?摆在统计公平进路面前的选择似乎是要么放弃其中的某一种定义,要么寻找这些定义之间的某种微妙平衡[13]。无论如何,统计公平都处于一种极为两难的境地。

### 2.2　公平悖谬困境

“辛普森悖论(Simpson's Paradox)”[16]是统计学中长期探讨的一种现象,指的是在不同的子群体中观察到的趋势或结果,在汇聚到整个总体时可能发生反转或消失。该现象最为人所熟知的现实案例之一要属 1973 年的加州伯克利研究生录取歧视事件[17]。根据该事件的统计数据,从总体层面看,女性申请者的录取率(34.6%)要低出男性申请者(44.3%)不少,因此舆论普遍认为伯克利对女性申请者存在歧视。然而,由于录取工作是按院系开展的,当数据划分到不同院系的子群体时,竟发现各院系女性申请者的录取率反而却略高于男性申请者,也就是“歧视”反转了!同一组数据从不同角度看,却得出了截然相反的结论,那么其中究竟是否存在歧视?我们应该相信哪一组数据?

如果把此例中的伯克利想象成一个用于决定录取结果的算法程序,当算法的决策结果出现辛普森悖论的情境时,是否有某一种统计公平定义可以准确地衡量该算法的公平性?笔者认为没有。在此例中,对于申请者的实际结果标签是未知的,也就是只能从不带实际结果标签的统计均等定义来衡量公平性。而面对同一组数据的不同分类,统计均等也只能得出两个互相矛盾的悖谬结论,无法判断真正的公平性,这相当于该公平标准的失效。

### 2.3　实际标签偏见困境

“实际标签偏见”是指实际结果标签 $Y$ 中已经包含了偏见或歧视的情况。其实,实际结果标签所包含的偏见往往都是历史上社会不公的反映。众所周知,美国有着长期的黑人种族歧视的历史,这种历史性的不公对黑人群体造成了深远的负面影响,使其在经济和教育领域长期处于劣势,进而可能增加犯罪风险。这反映在统计数据中便是黑人罪犯的实际犯罪率要大于白人罪犯。如果不同群体实际结果标签的差异很大一部分原因是来自于历史性的歧视与不公,那么根据这些数据训练出的算法又有何公平可言?

统计公平进路中的很多定义都依赖实际结果标

② 基本率(又称流行率 prevalence)是指特定群体中实际得到某种结果的比例。比如 $P(Y=1|A=0)$ 可以表示黑人群体的实际的再犯率。若 $P(Y=1|A=0)\neq P(Y=1|A=1)$,则说明两个群体的基本率不同。

签 $Y$,而一旦 $Y$ 本身包含偏见,算法便会调整决策结果来“适配”这些偏见,从而加剧不公。一个典型的例子来源于美国警察的执法偏见,实证证据表明,警察在决定拘留和搜查哪些人以及巡逻哪些社区时,无论是隐含地还是明确地,都会考虑种族[18]。如果警察尤为关注黑人群体所在的社区,那么犯罪记录中黑人群体的犯罪率相比白人就会更高,而这很大程度上仅仅是因为警察很少去白人社区执法。当算法使用这些带偏见的实际数据来训练用于预测罪犯的算法时,也会理所当然地认为黑人种族的真实犯罪率更高,自然也会造成预测结果的偏见。这种情况下,即使强制算法结果满足某种统计公平定义的约束,如机会均等,那么为了两种群体的真阳率相近,可能就会降低白人群体的预测阈值,使更多无辜者被误判为罪犯。也就是说,这类统计公平定义虽然以观察到的实际标签为条件,但却无法判断标签本身是否带有偏见,即使知晓存在偏见,也无法仅通过统计手段去除。这使得统计公平定义在很多场景中还会加剧历史的偏见因素。

综上所述,“定义非兼容性困境”指出某些统计公平定义之间无法同时满足,从而难以抉择或权衡;“公平悖谬困境”指出统计公平定义在面临诸如辛普森悖论的情境时,会得到关于公平性的矛盾结论,导致统计公平标准失效;“实际标签偏见困境”指出如果实际标签隐含历史偏见,那么依赖实际结果标签的统计公平定义可能会加剧这些偏见。

除了这三大困境之外,上一节提及的关于统计公平定义所共有的后两个特点也说明了该进路的局限性。其一是只能处理群体层面的公平,而造成个体层面公平性的缺失。一方面,即使不同群体达成了某一指标下的公平性,群体中的一些个体也很有可能遭到不公的待遇。另一方面,当算法面对只有极少量个体的决策场景时,统计意义上的各种公平定义可能会失去意义。因此,我们需要群体公平性的同时,也需要个体公平性的衡量标准;其二是只关注算法结果本身,而不关注算法何以产生结果。统计公平进路总是基于数据之间的统计关联性做公平判断,而不考虑数据之间更高层次的因果关联性。笔者认为,这一特点与以上三大困境的产生有很紧密的关系,解决困境的思路或许可以从探求数据之间的因果关系入手。

## 3 因果公平进路及其对困境的回应

实际上,算法公平与否归根到底是要判断受保护属性是否“引起”了算法结果的改变。“引起”是因果研究中的一个核心概念,若仅用统计学范式的方法处理,是否足够?计算机科学家、“图灵奖”获得者朱迪亚·珀尔(Judea Pearl)以他长期对统计学和因果理论的探究,告诉了我们一个确定的否定回答。珀尔认为,当前大多数机器学习研究者仍主要基于“数据为中心(数据拟合)”的统计学范式,他们认为机器理性决策的秘密来源于数据本身。然而当前越来越多的实践表明,仅靠数据本身无法产生智能决策所需的理解能力,而需要构建数据何以生成的因果解释机制,即从“数据拟合”转向“数据解释”范式[19]。也就是说,从统计公平进路出发,我们无法真正回答受保护属性究竟是否对算法的预测结果有所影响?也无法回答它们之间的影响是间接还是直接?更不清楚其他关键变量的影响机制如何。因此,要回答这些问题,我们必须从数据拟合范式转向数据解释的因果范式。很多算法公平的研究者意识到了这一点,从而开创了一条因果进路的算法公平性研究。

### 3.1 因果公平进路

因果公平进路的核心思路是首先揭示统计数据中隐含的因果关系,再利用因果推理方法检测受保护属性是否对算法预测结果产生因果影响,从而衡量算法的公平性。该进路主要依赖珀尔的结构因果模型(Structural Causal Model)开展因果推理③。

结构因果模型是一种形式化表征因果关系的模型方法,由一系列变量和变量之间的结构方程组成。变量是具体场景中我们要考察的对象,而变量的结构方程便是定量层面对因果影响机制的刻画。每一个结构因果模型都对应一个因果图(有向无环图),因果图可以看作是变量间定性层面因果关系的直观刻画。我们可以简单地用符号和箭头的形式来表达某一因果图结构,比如“$A\rightarrow X\rightarrow Y$”代表了从 $A$ 到 $X$ 再到 $Y$ 的一个因果链条。然后,珀尔定义了一个极为重要的形式算子——*do* 算子,用来表达对模型中

③ 结构因果模型的形式化定义可参见[19]。

变量的“理想化干预(ideal intervention)”[19]。比如,$do(X=x)$就是对变量 $X$ 进行干预使其值固定为 $x$。这种理想化干预可以理解为对该“原因”变量 $X$ 施加了一种完全无混杂,且不影响其他变量的假设性控制,以此来判断某一“结果”变量 $Y$ 有何相应的改变。如果对变量 $X$ 在 $do$ 算子的干预下,变量 $Y$ 的值或概率分布产生了改变,那么我们就把这种改变称为 $X$ 对 $Y$ 产生的“因果效应(causal effect)”。

也就是说,借助结构因果模型和 $do$ 算子的干预,我们可以判断两个变量之间有怎样的因果影响。因果公平进路正是抓住了这一点。如果我们对受保护属性进行不同值的干预,算法的预测结果或其概率保持不变,那么自然就说明受保护属性并没有对算法结果有影响,公平性就得到了确保。

因此,因果公平进路的第一种公平性定义便是基于这一思想,称为“纯干预公平(purely interventional fairness)”[11],需要满足以下等式:

$$P(\hat{Y}=\hat{y}\mid do(A=0))=P(\hat{Y}=\hat{y}\mid do(A=1))。$$

值得注意的是,从定义的形式来看,纯干预公平与统计均等之间的区别只在于多了“$do$ 算子”,然而它们的含义是截然不同的。珀尔曾将因果关系划分为三个不同的层级,即著名的“因果关系之梯”[20],第一层级是“观察”所得的统计关联性(非因果),第二层级是“干预”所得的群体层面(普型)的因果关系,第三层级是“反事实”所得的个体层面(殊型)的因果关系。统计均等定义只涉及观察所得的条件概率(统计关联性),而纯干预公平则通过干预算子“攀登”上了第二层级的因果。

纯干预公平衡量的是群体层面的公平性,因果公平进路的另一种定义正好弥补了个体层面公平性的缺失,即“反事实公平(counterfactual fairness)”[21]。反事实公平处理的是具体情境中个体之间的公平性,对应于“因果之梯”的第三层级。之所以称其为“反事实”,是因为其考虑的是不同于现实已发生情况(事实)的可能情境。以某招聘场景为例,将某求职者的属性分为两类,一类是受保护属性 $A$(如性别),另一类是包括学历、工作经验在内的其余重要属性 $X$。当某一女性求职者没有被成功录用后,她可能会考虑:假如我是男性,同时保持其余属性不变,那么录用结果会有改变吗?如果答案是肯定的,那么说明该招聘存在对个体的性别歧视,反之则说明并不存在对该个体的性别歧视。

反事实公平定义需要满足以下等式:

$$P(\hat{Y}_{A=0}(U)=\hat{y}\mid X=x,A=0)=P(\hat{Y}_{A=1}(U)=\hat{y}\mid X=x,A=1)。$$

上式中的 $U$ 是用来确定每一个个体的背景变量,作为条件出现的 $X$ 和 $A$ 是该个体具有的事实属性,而预测标记 $\hat{Y}$ 的下标“$A=0$”和“$A=1$”即代表对受保护属性 $A$ 进行了反事实层面的干预④。由于等式左边作为下标的干预“$A=0$”和事实“$A=0$”是一致的,因此其代表的是现实情境;等式右边的下标干预“$A=1$”与事实不一致,因此其代表的是反事实情境。如果反事实情境下的预测结果与现实情境的预测结果相等,那么就说明满足了反事实公平。

### 3.2　因果公平进路对统计公平进路面临困境的消解

首先,因果公平进路中的定义不会因为不兼容性而造成困境,相反还可为统计公平定义的不兼容性提供出路。上一节已经说明,当不同群体的基本率不同时,预测均等无法和假阳率或假阴率均等同时满足。而基本率的不同表明受保护属性 $A$ 与实际结果 $Y$ 具有某种相关性。反事实公平的提出者库斯纳(Matt Kusner)等人认为,当 $A$ 与 $Y$ 是相关的,这几种统计定义都不足以衡量公平性[21]。库斯纳指出,我们不妨假设 $A$ 是 $Y$ 的直接原因,由于 $A$ 直接影响 $Y$,因此算法需要确保不以实际结果 $Y$ 作为决策基础,否则算法预测结果 $\hat{Y}$ 必然受 $A$ 影响。利用结构因果模型,我们可以找到一些不被 $A$ 影响但组合起来能预测 $Y$ 的变量,根据这些变量所构成的函数去做算法决策,就能将受保护属性 $A$ 的影响去除。这样一来,统计公平定义之间的不兼容性就无关紧要了,因为 $A$ 与算法输出结果是独立的。

类似的思路还可以处理“实际标签偏见困境”。实际结果标签之所以有偏见,就是因为受保护属性 $A$ 通过某种直接或间接的方式影响了 $Y$。如果仍直接按照统计公平进路的方式,基于有偏见的实际结果标签来训练或衡量算法公平,当然还会固化偏见。更关键的点在于,统计公平进路甚至无法确定受保护属性是否对算法结果有直接或间接的因果影

④　珀尔将第三层级的反事实干预的 $do$ 算子省略,并作为结果变量的下标出现,此举是为了区分因果的第二层级和第三层级,前者表示群体层面的干预,后者则是个体层面的反事实干预。有关于此的具体讨论可参见[22]。

响[23]。而因果公平进路会对数据进行因果分析,构建出合理的因果假设,从而优先尝试在算法机制内部去除受保护属性对算法结果的影响。例如,上述库斯纳提出的方法便是这一思路。另一方面,从衡量算法结果公平性的角度看,因果公平进路通过对受保护变量的 *do* 算子干预,可以判断出受保护属性是否对算法结果有因果效应。若不同群体的因果效应有较大差异,则说明存在不公。归根到底,统计公平进路依赖实际结果标签来定义公平性,却无法处理其中的偏见,而因果公平进路不仅不依赖实际结果标签来定义公平性,还可以分析和去除其中的偏见。因此,因果公平进路自然不受困于实际结果标签的偏见。

最后,因果公平进路可以较好地应对辛普森悖论所造成的公平悖谬。珀尔通过对辛普森悖论的深入研究得出结论,只有理解统计数据背后的因果关系才能真正解决辛普森悖论[24]。对伯克利录取事件的统计数据进行因果分析后,人们发现女性更倾向于申请竞争更为激烈的院系,从而导致在总体上较低的录取率。也就是说,这其中存在一条"性别→院系选择→录取"的因果链条。由于院系选择完全出于个人意愿,我们不能把这一因果链条当做批判学校不公的依据。而只有当数据中存在"性别→录取"的直接因果影响时,才可认为有歧视倾向。因此,此例中的受保护属性"性别"可以允许通过"院系选择"来间接影响"录取",但不能直接影响"录取"。因果公平进路为这种受保护属性的"合理影响"提供了处理方案,因为利用 *do* 算子可以计算出"性别"到"录取"的直接因果效应,排除了"院系选择"的间接影响。那么,我们也就不必纠结于统计数据层面的公平悖谬了。更进一步,由于受保护属性从不同的因果路径可能具有不同的公平性影响,有学者提出了一套"特定路径公平(path - specific fairness)"[25]的方法来去除不公平路径影响的同时保留公平路径的影响。这样看来,受保护属性并不是绝对不能影响结果,而是要根据具体场景的因果假设做具体分析。

统观因果公平进路对三大困境的解决思路,根本上就是要揭示统计数据背后的因果关联。统计公平进路还只是停留在观察所得的"第一层级",由此产生了这些困境。此外,因果公平进路凭借衡量群体层面的纯干预公平和个体层面的反事实公平,很好地完善了统计公平进路只能处理群体公平的缺陷。而且,因果公平进路的这两种定义都基于结构因果模型这一较为成熟的理论框架,它们之间既有明显区别,又能互补统合。

## 4 余论:实现算法公平的方法论反思

从更高的角度看,统计公平进路与因果公平进路分别映射了两种不同的公平方法论。密西根大学的算法伦理研究者本·格林(Ben Green)区分了两种不同的算法公平方法论,即"形式(formal)算法公平"与"实质(substantive)算法公平"[13]。形式算法公平的方法论是指一种狭窄的公平性分析框架,该框架仅把公平性考量局限在孤立的算法决策程序中,也就是只根据算法本身的输入和输出来衡量公平性。格林认为形式算法公平至少会导致三点问题:第一,公平性只是算法技术层面的属性;第二,忽略算法上游的社会不公与偏见,比如历史数据的偏见;第三,无视算法下游对不公的固化。而实质算法公平作为一种具有更广阔视角的方法论,除了考虑算法本身决策的公平性,还要先确定上游数据中不同群体的差异是否隐含了不公,同时提出具体政策来减轻这些不公,另外还需考察算法能否促进具体政策的落实。格林的核心想法在于,算法公平性不应该只是简单地把社会中已实际存在的群体差异用某种公平定义给"看似"公平地表达出来,真正意义上的算法公平要为纠正社会中已有的历史偏见而努力。形式算法公平的目标只是前者,这便是"实质"与"形式"的主要差别。

根据格林对形式算法公平与实质算法公平的区分,本文发现,统计公平进路更多只是在"形式"上对算法的决策结果进行某种公平性约束或是衡量,其反映的就是形式算法公平所代表的狭窄分析框架;而因果公平进路的视域则更为宽广,该进路的因果分析框架不仅可以判断上游数据中是否存在偏见并予以处理,而且可以辅助制定改变这些偏见的实际政策,甚至还可为算法本身对于实际政策的落实是否有效提供帮助。因此,格林的这一方法论洞见与因果公平进路十分适配,统合起来可为算法公平研究开拓更多思路。

总的来说,因果公平进路提供了一个从因果角度分析公平性的良好框架,不仅可以回应统计公平进路所面临的困境,而且兼顾群体与个体层面的公平性。从当前人工智能发展的大背景来看,"黑箱"

算法的可解释性问题愈发重要[26]。如果我们无法给出算法何以公平或不公的解释,那么又该如何对算法的决策有充足的信赖? 因果公平进路的本质是通过因果模型的建构来增强模型决策的可解释性,从而使其更具透明性和可信赖性。其最终目标可以说不是追求绝对的"无偏模型",而是通过透明化的因果追溯,使算法的偏见生成机制可解释、可问责。因此,在对人工智能的可解释性诉求背景下,算法公平性研究从统计公平转向因果公平,是极为重要的发展趋向。

当然,本文并不认为因果公平进路应该完全取代统计公平进路,前者也仍然面临着诸多问题与挑战。比如具体因果模型的构建需要验证与评估,糟糕的模型反而会加剧歧视[27]。统计公平进路虽然存在着一些根本局限性,但它在某些场景也有其特有的评估和应用价值,因为某些统计公平定义更能反映出哲学中关于公平性的伦理观点,比如罗尔斯(John Bordley Rawls)的公平正义理论[28]。虽然格林区分了两种算法公平方法论,但他认为实质算法公平并不是对形式算法公平的取代,而是方法论的拓展[13]。本文同样也认为,因果公平进路也是在统计公平进路基础上的迈进,这一迈进有望使人工智能算法登上最高层级的"公平之梯"。

**参考文献:**


[1] Angwin J, Larson J, Mattu S, et al. Machine bias[EB/OL]. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. 2025-02-12.

[2] Duke T. Building Responsible AI Algorithms: A Framework for Transparency, Fairness, Safety, Privacy, and Robustness[M]. London, UK: APress, 2023: 61.

[3] Xu Z, Li J, Yao Q, et al. Addressing fairness issues in deep learning-based medical image analysis: A systematic review[J]. npj Digital Medicine, 2024, 7(1): 286.

[4] Saxena N A, Huang K, DeFilippis E, et al. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness[C]. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019: 99-106.

[5] 刘文炎,沈楚云,王祥丰,等.可信机器学习的公平性综述[J].软件学报,2021,32(5):1404-1426. Liu W Y, Shen C Y, Wang X F, et al. Survey on fairness in trustworthy machine learning[J]. Ruan Jian Xue Bao/Journal of Software, 2021,32(5):1404-1426

[6] Calders T, Kamiran F, Pechenizkiy M. Building Classifiers with Independency Constraints[C]. 2009 IEEE International Conference on Data Mining Workshops. IEEE, 2009: 13-18.

[7] Feldman M, Friedler S A, Moeller J, et al. Certifying and Removing Disparate Impact[C]. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015: 259-268.

[8] Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning[C]. Proceedings of the International Conference on Advances in Neural Information Processing Systems, 2016: 3315-3323.

[9] Berk R, Heidari H, Jabbari S, et al. Fairness in criminal justice risk assessments: The state of the art[J]. Sociological Methods & Research, 2021, 50(1): 3-44.

[10] Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments[J]. Big data, 2017, 5(2): 153-163.

[11] Loftus J R, Russell C, Kusner M J, et al. Causal reasoning for algorithmic fairness[J]. arXiv preprint arXiv:1805.05859, 2018.

[12] Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores[J]. arXiv preprint arXiv:1609.05807, 2016.

[13] Green B. Escaping the impossibility of fairness: From formal to substantive algorithmic fairness[J]. Philosophy & Technology, 2022, 35(4): 90.

[14] Larson J, Mattu S, Kirchner L, et al. How we analyzed the COMPAS recidivism Algorithm[EB/OL]. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. 2025-02-12.

[15] Dieterich W, Mendoza C, Brennan T. COMPAS risk scales: Demonstrating accuracy equity and predictive parity[J]. Northpointe Inc, 2016, 7(4): 1-36.

[16] Simpson E H. The interpretation of interaction in contingency tables[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1951, 13(2): 238-241.

[17] Bickel P J, Hammel E A, O'Connell J W. Sex bias in graduate admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation[J]. Science, 1975, 187(4175): 398-404.

[18] Lum K, Isaac W. To predict and serve?[J]. Signifi-

cance, 2016, 13(5): 14 – 19.
[19] 朱迪亚·珀尔.因果论:模型、推理和推断[M]. 北京:机械工业出版社,2022:Ⅳ – Ⅴ. Pearl J. Causality: Models, Reasoning and Inference, Second Edition [M]. Beijing: China Machine Press, 2022: Ⅳ – Ⅴ.
[20] 朱迪亚·珀尔,达纳·麦肯齐.为什么:关于因果关系的新科学[M]. 北京:中信出版社,2019:8. Pearl J, Mackenzie D. The Book of Why:The New Science of Cause and Effect[M]. Beijing: Citic Press Corporation, 2019: 8.
[21] Kusner M J, Loftus J, Russell C, et al. Counterfactual Fairness[C]. Proceedings of the International Conference on Advances in Neural Information Processing Systems, 2017: 4066 – 4076.
[22] 胡嘉伟.结构因果模型中的 do 算子分析[J].自然辩证法通讯,2024, 46(6): 57 – 65. Hu J W. An analysis of do – operator in structural causal model[J]. Journal of Dialectics of Nature, 2024, 46(6): 57 – 65.
[23] Kilbertus N, Rojas Carulla M, Parascandolo G, et al. Avoiding Discrimination through Causal Reasoning[C]. Proceedings of the International Conference on Advances in Neural Information Processing Systems, 2017: 656 – 666.
[24] Pearl J. Comment: Understanding Simpson´s Paradox [J]. The American Statistician, 2014, 68(1): 8 – 13.
[25] Chiappa S. Path – specific Counterfactual Fairness[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 7801 – 7808.
[26] 顿新国.可解释人工智能问题[J].江苏行政学院学报, 2023(3): 44 – 51. Dun X G. Problem with explainable Artificial Intelligence[J]. The Journal of Jiangsu Administration Institute, 2023(3): 44 – 51.
[27] Kusner M J, Loftus J R. The long road to fairer algorithms[J]. Nature, 2020, 578(7793): 34 – 36.
[28] Heidari H, Loi M, Gummadi K P, et al. A Moral Framework for Understanding Fair ml through Economic models of Equality of Opportunity[C]. Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019: 181 – 190.

## The dilemma and path forward in defining algorithmic fairness: From statistical fairness to causal fairness

HU Jia – wei
(School of Philosophy, Nanjing University, Nanjing 210023, China)

**Abstract**: The importance of artificial intelligence algorithms in various decision – making scenarios has become increasingly prominent, yet they simultaneously face challenges regarding fairness in decision – making. Currently, the dominant approach to defining algorithmic fairness is the statistical fairness approach, which measures fairness by setting specific statistical metrics. The core idea of the statistical fairness approach can be succinctly summarized as follows: if the algorithmic decision outcomes and protected attributes remain statistically independent under a given metric, then the algorithm satisfies fairness according to that metric. Generally, the definitions of fairness under the statistical fairness approach can be categorized into three types: statistical parity, statistical parity in accuracy, and calibration fairness. However, the statistical fairness approach faces at least three unresolved dilemmas: the "incompatibility of definitions dilemma," the "fairness paradox dilemma," and the "actual label bias dilemma." The "incompatibility of definitions dilemma" points out that certain statistical fairness definitions cannot be satisfied simultaneously, making it difficult to choose or balance between them. The "fairness paradox dilemma" highlights that statistical fairness definitions can yield contradictory conclusions about fairness in situations such as Simpson's paradox, rendering statistical fairness standards ineffective. The "actual label bias dilemma" indicates that if actual labels contain historical biases, statistical fairness definitions that rely on these actual outcome labels may perpetuate or even exacerbate these biases. This paper argues that these three dilemmas are constrained by the inherent limitations of the statistical paradigm and cannot be fundamentally resolved within the statistical fairness approach. The key to overcoming these challenges lies in shifting the paradigm for defining fairness. Inspired by Judea Pearl's causation theory, it becomes evident that the statistical fairness approach adopts a data – centric "statistical paradigm," which overly relies on data fitting while neglecting the causal explanatory mechanisms behind how the data is generated. From the perspective of the statistical fairness approach, it is impossible to truly determine whether protected attributes actually influence the algorithmic predictions, whether such influence is direct or indirect, or how other variables mediate this influence. Many researchers in algorithmic fairness have recognized this issue and have consequently pio-

neered a causal approach to algorithmic fairness. The core idea of the causal fairness approach is to first uncover the causal relationships within statistical data and then use causal inference methods to detect whether protected attributes have a causal effect on algorithmic predictions, thereby measuring algorithmic fairness. This approach primarily relies on Pearl's structural causal model for causal inference, mainly including "purely interventional fairness" and "counterfactual fairness." The causal fairness approach can effectively address the three dilemmas faced by the statistical fairness approach by revealing the causal relationships behind statistical data. By measuring purely interventional fairness at the group level and counterfactual fairness at the individual level, the causal fairness approach addresses the shortcomings of the statistical fairness approach, which can only handle group fairness. Moreover, both definitions under the causal fairness approach are based on the well – established theoretical framework of structural causal models, which are distinctly different yet complementary and unifiable. Finally, based on Ben Green's distinction between formal algorithmic fairness and substantive algorithmic fairness, this paper finds that the statistical fairness approach merely imposes certain fairness constraints or measurements on the algorithmic decision outcomes at a "formal" level, reflecting the narrow analytical framework represented by formal algorithmic fairness. In contrast, the causal fairness approach has a broader perspective. Its causal analytical framework can not only identify and address biases in upstream data but also assist in formulating practical policies to mitigate these biases. It can even help evaluate whether the algorithm itself effectively implements these policies. Green's methodological insight aligns well with the causal fairness approach, and their integration can open up more avenues for research in algorithmic fairness. Furthermore, the transition from statistical fairness to causal fairness reflects the core demands of explainability, transparency, and trustworthiness in artificial intelligence. Under these demands, the transition from statistical fairness to causal fairness represents an extremely important development trend in algorithmic fairness research.

**Key words**: algorithmic fairness; statistical fairness; causal fairness; structural causal model; methodology

---

(上接第 971 页)

applying the Relative Technological Comparative Advantage (RTCA) index to derive the Technology Gap Index (TGI), culminating in the algorithmic synthesis of these three cardinal metrics to yield the Comprehensive Technical Risk Value (TRV), which facilitates the stratification of technological risk severity into four discrete, ordinally ranked classifications: Extremely High Risk, High Risk, Medium Risk, and Low Risk; a rigorous empirical validation exercise concentrating on the strategically pivotal semiconductor domain—encompassing fabrication equipment, foundational materials, and finished integrated circuits—robustly corroborates the methodological feasibility and discriminatory power of the proposed system, with the application of the T – DCG model yielding the following granular risk categorizations: lithography machines, photolithography chemicals, logic processors and cutting – edge memory devices assessed at the Extremely High Risk level; radio frequency (RF) components, fourth – generation semiconductor substrates ion implantation systems, and photomask substrates classified as High Risk; silicon wafer substrates, atomic layer deposition (ALD) apparatus, plasma etching tools, and metrology/inspection equipment designated as Medium Risk; while chemical mechanical planarization (CMP) polishers, CMP slurry compounds, physical vapor deposition (PVD) sputtering targets, advanced packaging substrates, and lead frame components evaluated as Low Risk; this investigation consequently furnishes a novel theoretical vantage point for conceptualizing and dissecting technological risk, delivers an innovative, quantitatively grounded methodology for the ongoing evaluation and vigilant monitoring of technological vulnerabilities, and substantially augments the granular, micro – level scholarship pertaining to security risk appraisal within the intricate and dynamic realm of technology.

**Key words**: technology risk; assessment methodology; technology dependence; technology control; technology gap