系统性公平

Arindam Ray南佛罗里达大学，美国佛罗里达州坦帕，arindamray@usf.edu

Balaji Padmanabhan南佛罗里达大学，美国佛罗里达州坦帕，bp@usf.edu

Lina Bouayad
佛罗里达国际大学，美国佛罗里达州迈阿密，lbouayad@fiu.edu

arXiv:2304.06901v1 [cs.LG] 14 Apr 2023

摘要

机器学习算法日益广泛地应用于各种场景，用于制定或辅助决策。随着应用范围不断扩大，人们也越来越关注此类方法的公平性。以往关于算法公平的研究文献广泛探讨了相关风险，并在许多情况下提出了管理其中部分风险的方法。然而，大多数研究关注的是由某个（单一）核心决策者或主体采取的行动所引发的公平性问题。相比之下，大多数现实系统中有许多主体，他们作为更大生态系统的一部分协同工作。例如，在贷款场景中，有多个贷款机构为申请人评估贷款，还有政策制定者和其他机构，它们的决策也会影响结果。因此，单个决策者做出的任何贷款决策所产生的更广泛影响，很可能取决于生态系统中多个不同主体的行动。本文提出了企业公平性与系统性公平的形式化方法，并呼吁算法公平研究文献更加关注现实情境中整个生态系统的公平性，或者更简单地说系统性公平。

***K*eywords** 系统性公平 · 复杂系统 · 纵向动态 · 公平性模拟 · 长期公平性

1 引言

机器学习（ML）算法正日益应用于各种场景，从较为简单的应用（如推荐电影），到影响深远、可能实质性影响人们生活的应用（如帮助决定贷款申请的结果）。在许多此类场景中使用算法的优势显而易见。与人类决策者不同，算法能够处理大量数据、执行复杂计算，随着决策数量增加而扩展规模，而且不会感到疲倦。

人们还认为，算法决策更加客观、公平，因为它们不受人类主观性和偏见的影响。这一点很重要，因为机器学习算法被广泛用于多个领域的决策，并回答诸如面试者能否胜任某个职位 [1], 大学申请者能否取得成功并毕业 [2], 贷款申请者是否会违约 [3] 或者被捕者是否可能成为累犯 [4]等问题。机器学习驱动的预测显然正在影响重大决策；因此，确保这些预测及其伴随的决策不会无意中伤害某些个人群体至关重要。

然而，现实中我们发现了一些反例，对算法方法能够做到公平这一盲目假设提出了质疑。美国法院用于作出假释决定的工具“矫正罪犯管理及替代制裁分析系统”（COMPAS）被发现，在画像相同的情况下，它更有可能给非裔美国罪犯评定较高的风险评分，而对白人则不会 [5]。亚马逊放弃了一款招聘工具，因为所用算法歧视女性候选人 [6]。在美国抵押贷款市场，情况相似的拉丁裔和非裔美国借款人从传统贷款机构借款时，利率会高出 7.9 个基点。更多依赖算法的金融科技平台表现稍好，但正如我们最近看到的，它们仍会多收 5.3 个基点 [7]，令人遗憾。

受这些问题的启发，研究文献中对如何更好地解决算法公平问题[8, 9, 10, 11, 12, 13, 14, 15, 16, 17]的兴趣日益浓厚。大多数公平研究文献聚焦于静态（也称为“单次”或“批处理”）设置 [11, 18] 下的分类问题，在这些设置中，完整数据可预先获取。此类基于真实数据集的评估确实具有贴近实际的优势。然而，正如 [19] 所论证的那样，存在一些长期公平性问题，无法通过这类静态设置加以理解和解决。

为了评估单次或静态设置之外对公平性的影响，一些研究 [20, 21, 22]采用了两阶段模型。在这些模型中，第一阶段的决策作用于来自两个不同人口群体的个体，这可能促使个体采取特定行动。随后在第二阶段分析这一单步干预对整个群体的总体影响。此外，还有研究从分析角度探讨两阶段以上动态设置中的公平性，涉及大学招生 [23, 24],、招聘 [23],、贷款 [25],、劳动力市场 [26],、文本自动补全 [27]和语音识别 [28]。这些研究表明，决策系统中的公平性考量需要理解其对相关人口的影响；否则，看似公平的约束实际上可能会加剧人口中各群体之间已有的差异 [24]。

在以现有文献为基础研究公平性的长期动态时，我们也注意到，目前大多数研究都是从单一决策者的角度考察公平性，例如由一所大学做出招生决策、一个组织做出招聘决策，或一家银行做出贷款决策。然而，在现实世界中，这些机构都不是孤立运作的。同一行业中的其他机构往往会观察某个组织的行动及其结果，进而采取应对决策，从而影响最终形成的结果。因此，我们需要深入理解系统性公平 的视角，而不应仅仅关注个体公平。即使从政策制定者的角度来看，重要的往往也是整个系统的行为，而非单个企业的行为。企业层面的公平性是否总是与系统层面的公平性相一致？我们又该如何研究这类问题？本文在 [19]的研究基础上，提出了一种研究方法：通过仿真考察系统层面的行为，并以贷款决策为例进行分析。

受这些考量的启发，本文研究了先前工作 [19] 中简化贷款场景的一个变体，主要旨在阐释若干关键观点，并论证文献中更广泛地研究系统性公平的必要性。具体而言，本文作出以下贡献：

1. 我们提出并研究系统性公平，这是对公平研究文献的一项新颖且重要的拓展。2. 我们将长期算法公平结果的研究拓展至多智能体系统。尽管已有研究探讨长期结果，但目前这些研究聚焦于单个主体 [19] ，且未讨论系统性公平与个体公平之间的区别。3. 我们开发了用于研究个体公平与系统性公平的形式化方法，并提出了研究系统性公平的框架。4. 我们通过研究贷款示例中涉及两家企业的具体情景，给出研究结果，并以此阐明：除了个体公平之外，考虑系统性公平的差异与重要性。5. 在确立主要观点后，我们展示了如何借助关注理解涌现行为 及由此产生的系统性公平问题和结果的复杂系统视角，轻松扩展该框架，以便对多家企业和多个主体进行建模。

2 背景与文献综述

虽然我们重点关注算法公平性，但公平性这一概念早已在心理学、社会学和哲学领域得到研究，以考察偏见与歧视。为了面向广大读者，《剑桥词典》将公平性定义为“平等地对待他人，或以正确、合理的方式对待他人的质量”。然而，从原则上来说，对这一概念进行衡量相当复杂，尤其是在机器学习算法的背景下，因此也产生了许多关于公平性的不同定义。为了便于展开本文的讨论，这里简要列举其中一些定义，摘自 [29] 和 [30]。

ˆ 对于形式化定义，我们使用以下符号： Y 表示实际结果， Y 表示预测结果。我们假设结果是二元的——1表示有利结果，0表示其他情况。 $P \in \{1, 0\}$ 表示受保护属性，其取值可以是1或0，具体取决于个人是否属于受保护群体。

统计均等 [8] （也称为**群体公平性** [8], **人口统计均等** [29]）： 当一个人是否属于受保护群体不影响其被预测为正面结果的可能性时，即当 ˆ ˆ$P(Y = 1|P = 1) = P(Y = 1|P = 0)$ 时，便满足这一条件。

条件统计均等 [12]: 先前的定义忽略了可能导致不同群体之间结果差异的其他因素，因此，该定义考虑了一组可能影响结果的合法因素 L。根据这一定义，如果在控制一组合法因素 L 后，无论个人是否属于受保护群体，其获得正向预测结果的可能性都相同，那么该算法就是公平的，即 $P(\hat{Y} = 1|L = l,\ P = 1) = P(\hat{Y} = 1|L = l,\ P = 0)$。

预测均等 [12]（也称为**假阳性错误率平衡**）： 如果某人属于负类却被错误预测为正类的可能性.....，则分类器算法满足此定义。

无论个人是否属于受保护类别，该指标都保持不变，即 $P(\hat{Y}=1|Y=0,\ P=1) = \hat{P}(Y=1|Y=0,\ P=0)$。换言之，受保护群体和非受保护群体具有相同的假阳性率（FPR），因此该指标也称为假阳性错误率平衡。对于两个群体具有相同假阳性率的分类器，其两个群体的真阴性率（TNR）也将相同，即 $P(Y=0|Y=0,\ P=1) = P(Y=0|Y=0,\ P=0)$。

机会均等 [11]（也称为**真阳性错误率平衡**）：如果一个分类器算法满足以下条件，则认为其符合这一定义：无论一个人是否属于受保护类别，其被正确划分为正类的可能性都相同，即 $P(Y=1|Y=1,P=1) = P(Y=1|Y=1,P=0)$。换言之，受保护群体和非受保护群体具有相同的真正率（TPR），因此这一概念也称为真阳性错误率平衡。从数学角度来看，对于两个群体而言，具有相同 TPR 的分类器也将具有相同的假阴性率（FNR），即 $P(Y=0|Y=1,P=1) = P(Y=0|Y=1,P=0)$。在本文中，我们将采用这一定义来研究系统性公平，但需要注意的是，结果当然取决于所研究的定义。虽然我们也展示了采用其他定义时的结果，但我们强调，此处的重要之处在于概念化和框架，它们能够广泛推广到许多定义和应用场景。

以上公平性定义属于**群体公平性**的范畴，强调平等对待不同群体（受保护群体与未受保护群体）。另有一类公平性定义源于这样一种理念：为相似的个体提供相似的预测；这类定义被归为**个体公平性** [8, 13]。

已有大量研究探讨机器学习中的公平性。其中大多数研究采用真实数据集，因此与现实相关 [11, 18]。但近期研究 [19] 指出，这类研究在考察长期公平性方面能力有限；要研究长期公平性，必须能够收集经过精心整理的纵向数据并进行实验处理，或者通过仿真加以近似。表1总结了相关文献中的关键观点，重点关注公平性在较长时间尺度上的动态。

参考文献	时间范围	公平性概念	所用方法	领域
[20]	两阶段	差分隐私、机会均等	分析型	贷款
[21]	两阶段	EO	分析型	社会学习
[22]	两阶段	机会均等，其他	分析型	大学录取
[26]	长期	其他	分析型	劳动力市场
[27]	长期	其他	分析型	文本自动补全
[28]	长期	差分隐私，机会均等	分析型	语音识别
[24]	长期	DP	分析型	大学录取
[19]	长期	EO	强化学习	贷款、大学录取
[23]	长期	机会均等、预测均等	分析型	大学录取、招聘
[25]	长期	DP	分析型	贷款
本文	长期	*EO*	基于智能体的建模	贷款

表1：公平性相关研究汇总。注：差分隐私－人口统计均等，机会均等－机会均等，预测均等－预测均等 [30]

如表1所示，先前文献中的大多数研究采用分析型模型来评估算法公平性。在这些模型中，个体属于受保护群体或非受保护群体。每个个体都试图完成一项任务（例如申请工作或大学）。决策者会批准或拒绝个体的请求。作者利用分析型模型评估决策者的行为对个体长期产生的影响。例如，在劳动力市场中，基于分析型模型开展的长期研究表明，在实行统计均等招聘的情况下，即使不同群体最初的社会地位存在差异，经过一段时间后，各群体的就业前景和薪资前景也不会出现显著差异。然而，群体盲视方法和统计歧视性招聘（利用群体成员身份推断隐藏属性）会导致群体声誉分化[26]。

然而，所有这些研究的重点都是应用于单个焦点企业或决策者情境中的公平性。正如我们前面所指出的，此类方法并未将生态系统中多个实体行动所产生的涌现结果纳入考量。我们的研究补充了当前关于长期公平性的研究，并展示了建模公平性时超越单个决策者的必要性。所提出的框架可被视为仿真和分析方法的测试平台，用于测试假设情景及其涌现结果。

3 形式化定义

在本节中在本节中，我们定义本文用于研究系统性公平相关动态的关键术语。 s.

个体 - 个体是决策所针对的人。每个个体 k，其中 $k \in (Z) : k \in (0, n)$，都有一个质量评分 $Q_k \in [0, 1]$，该评分表示该个体 k 的成功概率。质量是不可观测的。然而，质量是可观测特征 X_k 的函数，而这些特征被用于做出决策。个体还具有一项受保护属性 $P \in \{0, 1\}$，用于表明其是否属于受保护类别。做出决策时不得使用受保护属性。

$$Q_k = \phi(X_k) \tag{1}$$

在贷款场景中，个体是寻求信贷的借款人。严格来说，借款人仅指申请并获得了信贷的人；为简洁起见，我们将采用上述定义。在此语境下，如果借款人的质量评分为 0.9，我们就将其理解为该借款人偿还贷款的概率为 0.9。

群体 - 群体是根据受保护属性划分的个体集合。群体可以是受保护群体（$P = 1$），也可以是不受保护群体（$P = 0$）。

企业 - 企业是在所考虑的情境中采取行动并作出重要分类决策的组织。在我们的贷款场景示例中，企业是向借款人发放贷款的贷款机构（银行）。

生态系统 - 生态系统是由为个体提供服务的 m 企业、个体集合，以及政策制定者等其他实体构成的集合；这些实体的行动会以各种方式影响贷款或还款决策。为便于说明，我们在此仅考虑个体和企业；然而，要在各种情境下建立现实的模型以理解系统性公平，就必须在复杂系统框架中恰当地对所有主体及其互动进行建模。

决策 - 每家企业 l 都有自己的算法，根据个体 k 的可观测特征 X_k来估计其质量。我们假设企业具有一个算法复杂度参数 s_l；复杂度水平越高，估计质量就越接近实际质量。

$$\hat{Q_{lk}} = f(X_k, s_l) \tag{2}$$

在对个人质量进行估计后，企业 l 针对个人 k 的决策取决于企业的风险阈值 τ_l，并且可以表示为

$$D_{lk} = \hat{Q_{lk}} > \tau_l \tag{3}$$

在贷款场景中，决策是批准还是拒绝贷款，而这实际上是在尝试预测申请人的质量（即他们是否会偿还贷款）。

结果- 结果 O 是指决策对个体产生的影响。我们将结果与决策区分开来，因为结果可能取决于一个或多个决策。

$$O_k = g(D_{1k}, D_{2k}, ..., D_{mk}) \tag{4}$$

在贷款场景中，个人按时偿还贷款后，其信用评分可能会提高。然而，如果贷款未获批准，信用评分则保持不变。另一方面，如果个人的贷款获得批准，但随后发生贷款违约，其信用评分就会降低。

决策公平性 - 第 2 节讨论了多种可用于此目的的公平性定义。不过，为了便于说明，我们主要采用<escape></*escape*>机会均等<*escape*></escape>指标来衡量决策公平性。具体而言，当实际值为正时，被预测为正（即 $D_l k == 1$）的人所占的比例，即真正例率（TPR），在受保护群体和非受保护群体中应当相同。决策公平性是在企业层面进行衡量的。我们通过分析$TPRGap$来研究这一问题，即受保护群体的真正例率与非受保护群体的真正例率之间的差值。

结果公平——与上述概念化类似，当实际值为正时，结果为正（即 O_k == 1）的人所占比例在受保护群体和非受保护群体中相同时，即体现了结果公平。为说明系统性公平，在此我们也将其视为**生态系统公平**，因为结果取决于生态系统中所有企业的决策。然而，正如公平性有多种定义方式，生态系统公平的衡量指标也可以有多种，这部分留待未来研究。

我们在 2×2 设置中考虑四种情景（图1）。在横轴上，我们考察时间范围——决策是一次性完成，还是属于长期场景，需要多次作出决策。在纵轴上，我们根据决策者是一个还是有多个实体作出类似决策来划分。

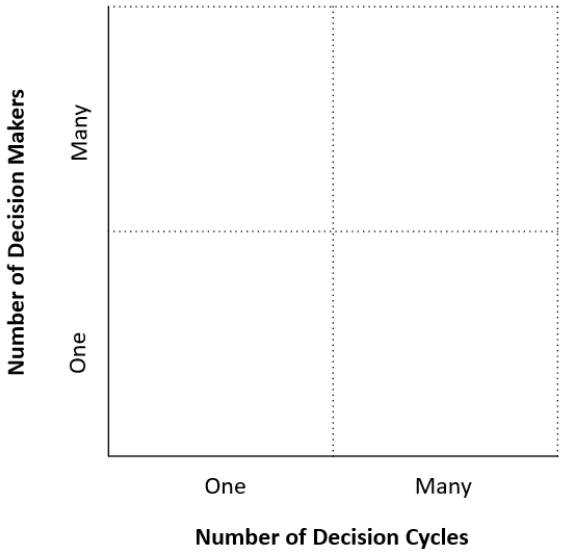


图1：算法决策情景

快照与长期场景之间的主要区别在于反馈的影响，即个体会因其成功而获得奖励，因其失败而受到惩罚。对于长期场景，我们进一步定义以下附加术语。

奖励 - 奖励， ΔQ^+ 定义为个人 k成功后，其 Q_k 的增加量。例如，在贷款场景中，个人偿还贷款。

惩罚 - 惩罚 ΔQ^- 定义为其未能成功时， Q_k的减少量。例如，在贷款场景中，个人未能偿还贷款，构成违约。

4 模拟研究 1

4.1 设置

为了评估我们的框架，我们实现了一个基于智能体的模拟。我们考虑一个由 1000 个个体组成的人口，他们以相同的概率属于受保护群体或非受保护群体 [1]。每个个体还具有介于 0 和 1 之间的质量评分。按照 [19]，中的设置，我们调整了受保护群体的质量评分分布，使其一开始处于不利地位——其分布与非受保护群体相同，但整体左移。每个时间段，都会随机选出 100 人，向生态系统中的企业申请贷款。

为使解释尽可能简单，我们在此考虑单企业场景，或模拟中包含两家企业的多企业场景（随后我们会放宽这一假设，并展示更大规模设定下的结果）。此外，我们用两个方面来对每家企业进行参数化：复杂度 和阈值。复杂度表示企业用于估计贷款申请人质量的机器学习算法的质量（即准确性或其他指标）。复杂度越高，申请人的真实质量评分与估计质量评分之间的误差就越低。阈值表示银行的风险偏好。给定对某个人质量评分的估计，企业批准该人的最低评分是多少?

由于我们的情境是个体申请贷款（可能是个人贷款、信用卡、汽车贷款等），并且允许同一个体重复申请（如现实生活中一样），因此我们选择了50个时间段，认为这足以涵盖个人生命中足够长的时间范围（这一时间范围也足以识别此处的主要定性结果）。根据机会均等的定义，我们在每次迭代结束时测量受保护群体和非受保护群体的真正例率。为了消除仿真模型随机性造成的影响，我们针对每组参数执行100次模拟，并取结果的平均值。

[1]我们理解，在现实中，受保护群体的代表性不足。不过，为便于阐述，我们先让模拟保持简单，不考虑类别不平衡带来的额外复杂性。尽管如此，我们已在附录 A 中针对受保护群体通常代表性不足的场景重复了模拟研究 1，并发现我们的观察结果在该场景下依然成立。

模拟 Parame-ter	价值观	描述	形式化链接
n	1000	生态系统中的个体数量	个体
f	0.5	受保护群体中的个体比例	群体
m	1, 2	生态系统中的企业数量	企业)
t	50	决策周期数	决策周期
τ_l	0.5、0.6、0.7、0.8	企业批准个体的质量阈值——ual	企业属性 – 风险容忍度-容忍度
s_l	0.7，0.9	实际质量与预期质量之间的噪声	企业属性 – 模型复杂度
ΔQ^+	+0.05	奖励——成功的个体的质量评分变化。得	奖励
ΔQ^-	-0.05	惩罚 - 未通过者个人的质量评分变化	惩罚

表2：仿真参数值汇总

对于 s 和 τ，尽管可能的取值范围是0到1，但我们采用了较小的取值范围，以更贴近实际情况。为确保全面性，我们针对 s 和 τ 的整个取值范围重复了实验，并将结果记录在附录C中。

为建立反馈机制，我们根据个人的真实结果对其进行奖励或惩罚，并将其反映在其质量评分中。为简单起见，我们将奖励和惩罚设为相同的幅度，使其质量评分变化为 $+0.05$ 或 -0.05 。表2总结了我们模拟中使用的仿真参数及其取值。

4.2 结果

使用我们的仿真模型，我们研究企业和生态系统属性对有效性和公平性的影响。我们设置焦点企业（FF）的模型复杂度固定的情景。随后，我们改变 FF 的审批阈值，并观察次级公司（SF）的不同设置对企业和生态系统公平水平的影响。在所有情景中，我们衡量：1）模型识别高质量借款人的有效性（真正例率）；以及 2）这些模型在企业和生态系统层面的公平水平（真正率差距）。

4.2.1 模型公平性

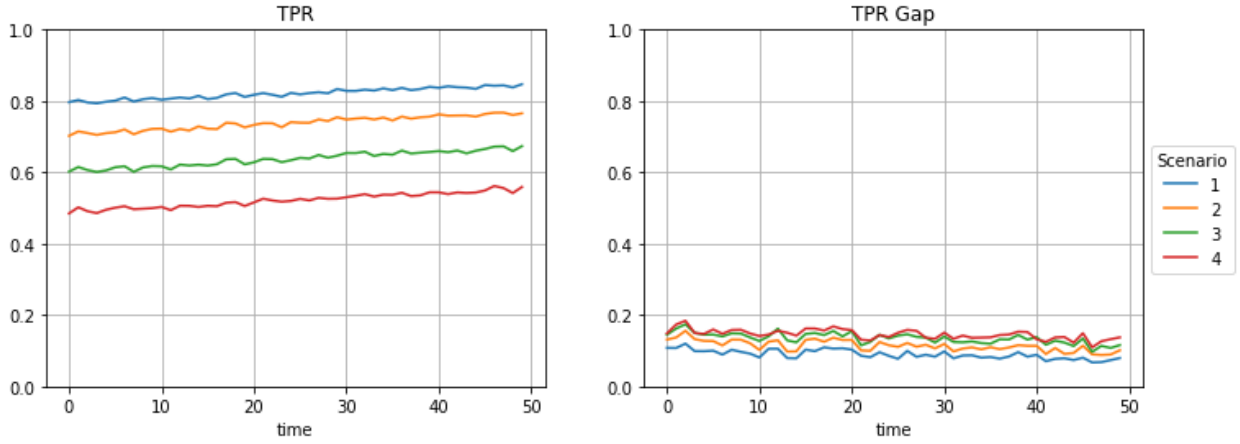
为探究企业属性对公平性的影响，我们考察了不同情境下受保护群体与非受保护群体之间的真正率差距。

考察模型复杂度相对较高的企业（图3），我们可以清楚地看到审批阈值对企业公平性水平的影响（图3a）。随着高度复杂的FF的审批阈值不断提高，该企业将把申请范围限制在质量非常高的借款人之中。鉴于受保护群体中个体的质量水平相对较低，这种设置将使非受保护群体中的个体获得更多优势。值得庆幸的是，在这一生态系统中增加第二家企业（图3b和图3c），可以抵消阈值水平的影响，并有助于提升公平性。第二家企业采用较低的阈值，使受保护群体中的个体有机会获得批准。这表明，在实践中，当政策制定者观察到审批阈值非常高时，可以鼓励市场竞争，甚至专门设立面向受保护群体中个体的机构贷款。

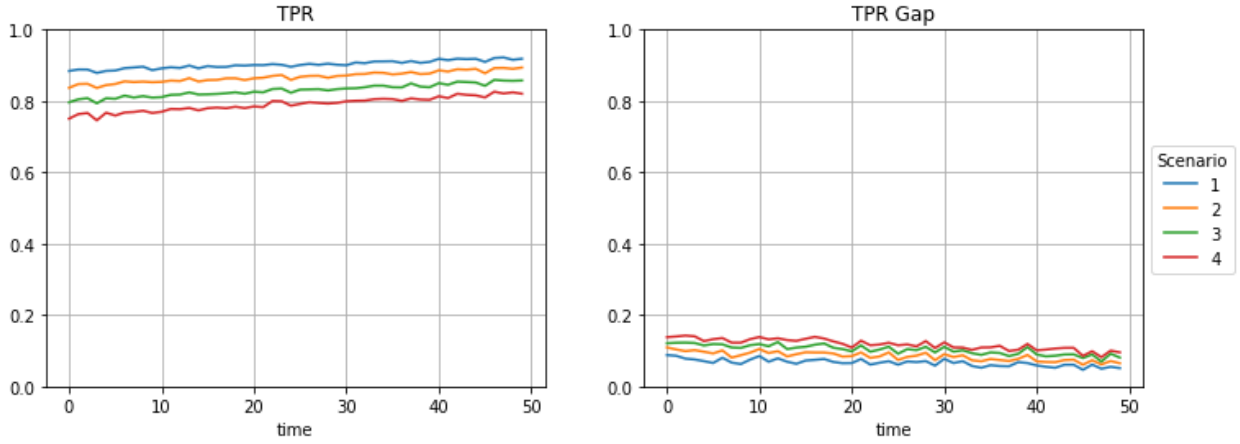
我们观察到，在单企业场景中，当复杂度参数保持不变时，阈值的增加会导致非受保护群体与受保护群体之间的真正率差距扩大（图2a和图3a）。随着阈值不断提高，企业批准的受保护群体成员数量减少，因此真正率差距也随之扩大。

我们在多企业场景中也观察到了相同的行为（图2b和图2c，以及图3b和图3c）。此外，与生态系统中的各企业单独运营时各自对应的真正率差距相比，多企业场景下两个群体的生态系统真正率差距均更低。换言之，随着我们向该市场中增加一家新企业，生态系统层面的公平性也会随之提高。另一个值得关注的观察结果是，随着审批阈值的提高，两个群体之间的真正率差距也会扩大。然而，与单企业场景下各组成企业的真正率差距相比，所有多企业场景中的这一差距均较小。

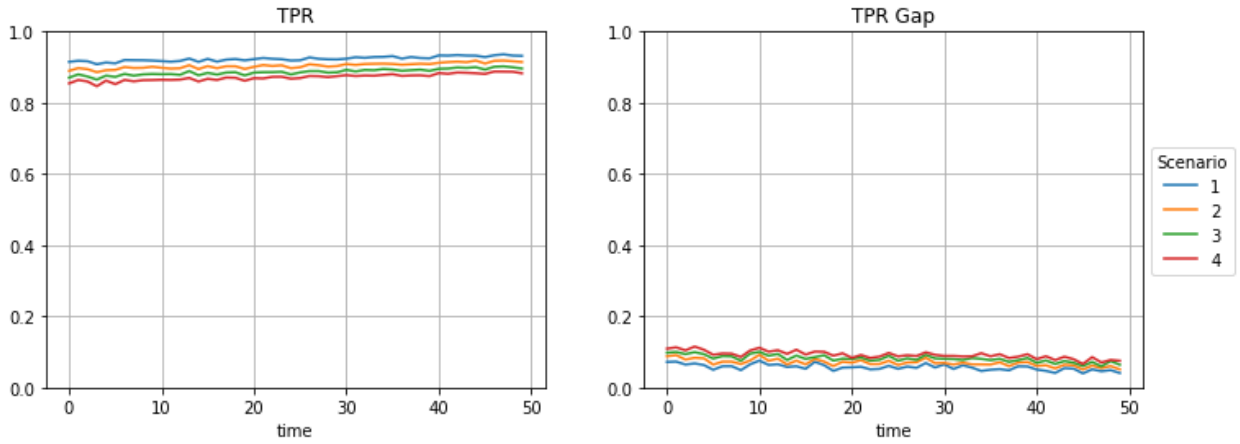
虽然这里我们的模拟研究侧重于将真正例率差距（均等机会）作为公平性度量，但我们表明，采用统计均等差距作为公平性度量时，得出的推论同样成立（附录B）。



(a) 单一企业、焦点企业 $S = 0.7$

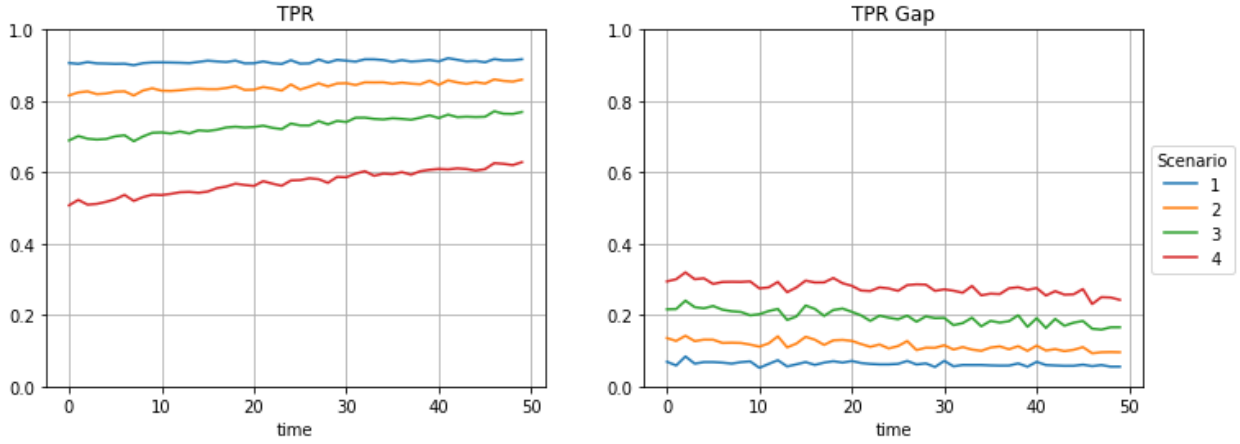


(b) 多企业、焦点企业 $S = 0.7$、第二家企业 $\tau = 0.7,\ s = 0.7$

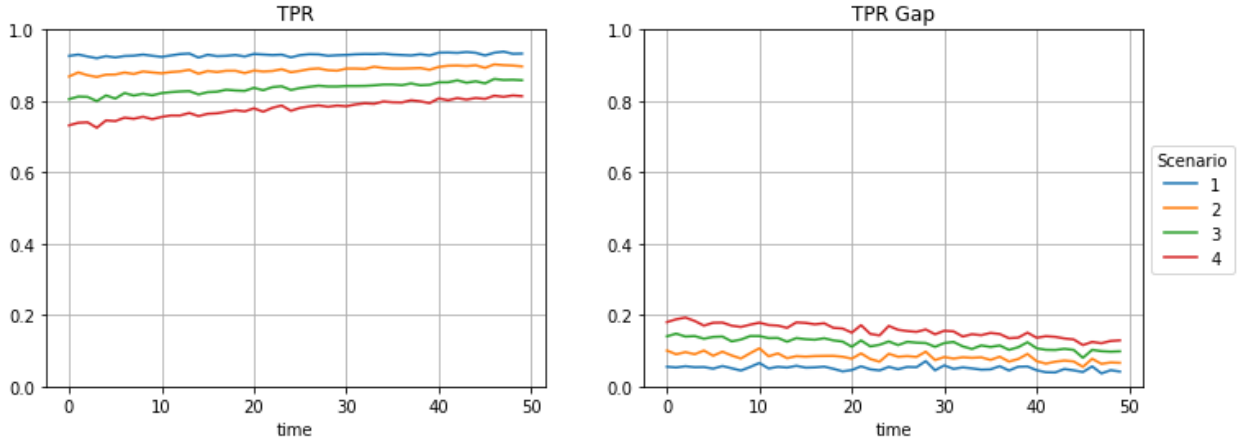


(c) 多企业、焦点企业 $S = 0.7$、第二家企业 $\tau = 0.6,\ s = 0.9$

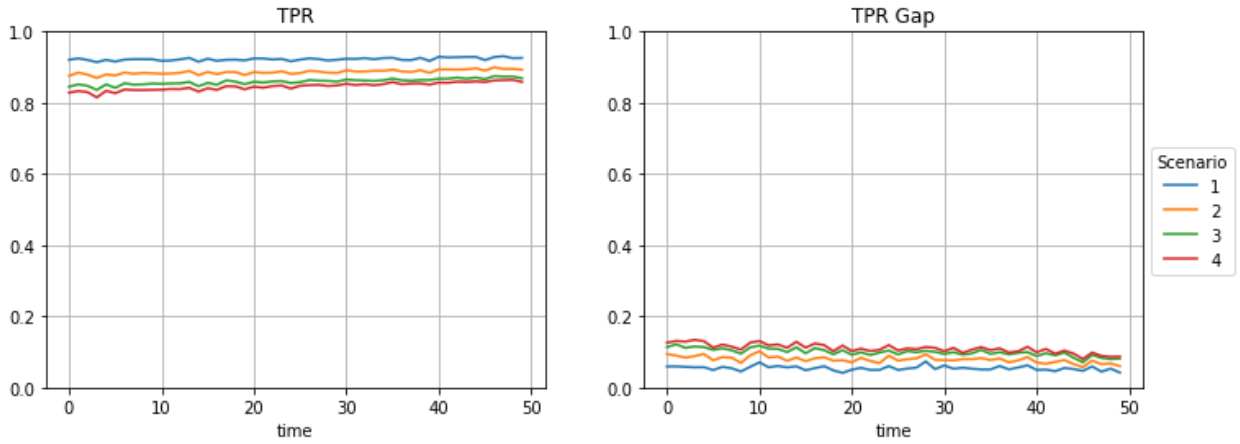
图2：在焦点企业复杂度较低的单一企业和多企业场景中，受保护群体与非受保护群体之间真正率及真正率差距的变化。在所有子图中，情景1、2、3和4分别表示焦点企业的阈值为0.5、0.6、0.7和0.8。



(a) 单一企业，焦点企业 $S = 0.9$



(b) 多企业，焦点企业 $S = 0.9$，第二家企业 $\tau = 0.7$， $s = 0.7$



(c) 多企业，焦点企业 $S = 0.9$，第二家企业 $\tau = 0.6$， $s = 0.9$

图3：焦点企业复杂度较高时，单一企业和多企业场景中受保护群体与非受保护群体的真正率及真正率差距的变化。在所有面板中，情景1、2、3和4分别表示焦点企业的阈值为0.5、0.6、0.7和0.8

4.2.2 模型有效性

维护或提升公平性至关重要，也是本研究的重点。然而，同样重要的是评估这些模型在筛选出实际会偿还贷款的消费者方面的有效性。为此，我们跟踪上述情景中模型的真正率（TPR）。

当企业在市场中作为唯一贷款方运营时，随着审批阈值的提高，其真正例率（TPR）会降低（图2a，情景1、2、3和4）。这主要是因为，随着企业提高阈值水平，其批准的申请人所对应的预期质量范围会变得更加狭窄。在模型复杂度较低的情况下，企业无法区分这一群体中质量较高和质量较低的申请人。

然而，当次级公司（图2b和图2c）加入这一生态系统时，我们观察到其对模型有效性产生了显著影响。总体而言，将复杂度相同或更高、且审批阈值更高的企业引入市场，会提高真正例率（TPR）值。这可以解释为：由于申请人需要经历不止一次申请流程（审批阈值相对较高），组合后的模型会作为集成模型发挥作用，从而更好地从两个群体中筛选并批准高质量申请人。
在实践中，这意味着，当观察到个别企业的模型未达到最优水平时，政策制定者应鼓励增加市场中的企业数量，因为这将提高受保护群体和非受保护群体中符合条件的个体获得审批的可能性。

4.2.3 纵向效应

在上述所有情景中，我们跟踪了多轮贷款申请过程中模型的公平性和有效性水平。这使我们能够研究其长期影响。在我们的模拟设置中，我们观察到，随着时间的推移，模型的有效性（真正例率，TPR）不断提高。我们还发现，真正率差距（TPR Gap）持续缩小，这表明模型的公平性随着时间推移而不断提高。这可以解释为：偿还贷款的借款人会因其积极行为而获得奖励，其质量评分随之提高，反之亦然。由于这些奖励和惩罚效应，高质量申请人与低质量申请人之间的差距会随着时间推移而不断扩大，从而使模型更容易识别并批准高质量申请人。
总体而言，我们可以得出结论：在上述研究的设置中，多企业场景不仅能够提供更高的真正例率（TPR），还能缩小真正率差距（TPR Gap）。这表明，在生态系统层面，两个群体受到的待遇更加平等。更广泛地说，这也体现了在此类重要场景中采用系统性公平视角的价值。

5 模拟研究 2

5.1 设置

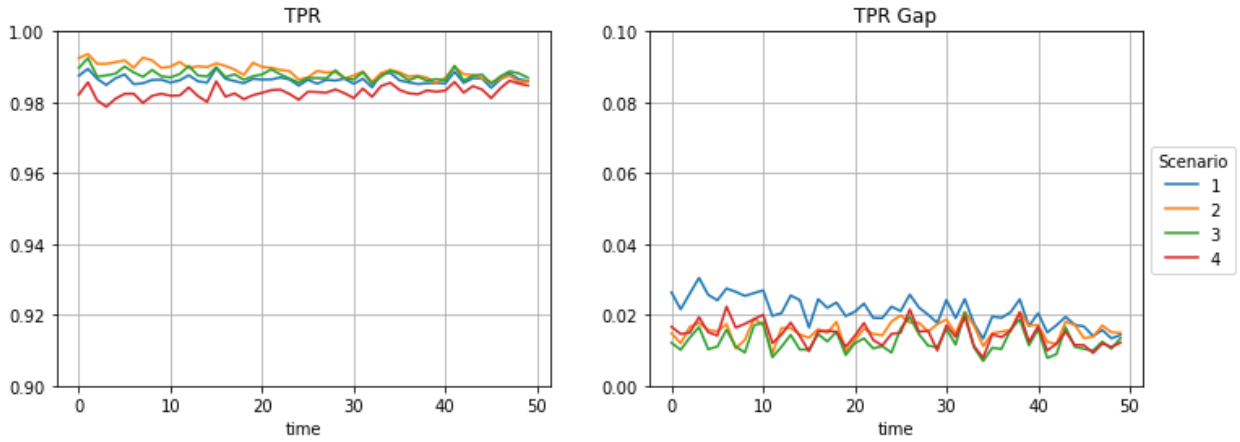
在前一项仿真研究中，我们假设企业针对受保护群体和非受保护群体具有相同的复杂度和阈值。本研究放宽了这些假设。

5.2 结果

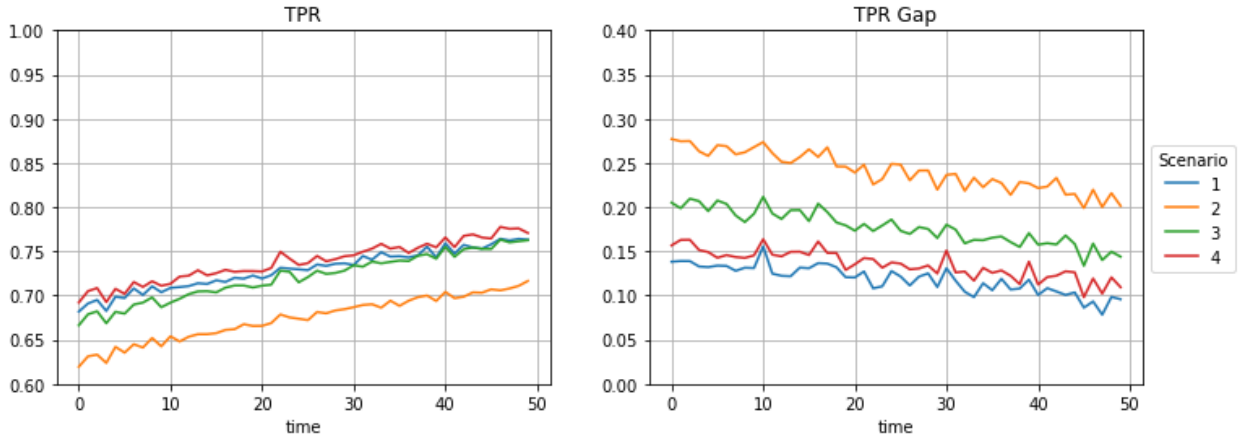
5.2.1 隐性偏见

在这种场景下，企业能够比受保护群体中的成员更好地评估非受保护群体中的成员。这可能是由于缺乏有关受保护群体的数据，也可能是因为此类数据中存在历史偏见。在我们的环境中，我们通过让企业对受保护群体和非受保护群体采用不同的复杂度水平来模拟这一情况。例如，在图4a的情景1中，与受保护群体（$s = 0.7$）相比，两家企业对非受保护群体采用了更高的复杂度（$s = 0.9$）。
然而，在受监管行业中，行业监管机构/政策制定者可能不接受这种场景，并可能建议企业将两个群体的复杂度水平设为相同。鉴于这项强制要求，企业可能采取三种行动：a）提高受保护群体的复杂度水平，使其与非受保护群体相匹配（图4a 情景2，$s = 0.9$ 两个群体均适用）；b）降低非受保护群体的复杂度水平，使其与受保护群体相匹配（图4a 情景4，$s = 0.7$ 两个群体均适用）；或 c）折中处理（图4a 情景3，$s = 0.8$ 两个群体均适用）。

在理想场景（如情景2所示）中，企业对两个群体都将具有更高的复杂度（$s = 0.9$），从而提高效率（更高的真正例率），同时改善公平性（更低的真正率差距）。然而，在现实中，实现这一目标可能更加困难，因此企业可能会选择场景3或场景4。在场景3和场景4中，公平性均有所改善（更低的真正率】【。



(a) 阈值较低（0.3）的企业



(b) 高阈值（0.8）的企业

图4：企业所表现出的隐性偏见的影响。

差距），而非场景中 1，且更接近情景2。效率、真正例率会随着复杂度降低而下降。情景4 i s 甚至比场景还要糟糕。情景1，其中至少非受保护群体的复杂度水平较高。

不过，上述结果会因企业的阈值水平而异。在图4b中，批准阈值较高（$\tau = 0.8$），此时较高的复杂度实际上会产生相反的效果。我们可以看到，在这种情况下，情景2的效率最低（真正例率最低），公平性也最差（真正率差距最大）。这一现象可以解释为：在这种情况下，由于阈值非常高，只有一小部分高质量人群获得批准。较低的复杂度允许质量估计存在更大的差异，因此更多质量仍然相当不错的人能够获得批准。因此，对于阈值较高的企业，降低复杂度会提升效率和公平性。

5.2.2 显性偏见

虽然我们认为这种做法在现实生活中可能并未实行，但我们仍从纯学术角度评估这一场景。我们的设计方式是：企业对受保护群体和非受保护群体设置不同的阈值。

我们观察图5中的结果。在情景1中，两家企业都是公平的，也就是说，两家企业对两个群体采用相同的阈值。在情景2中，企业1对受保护群体设置了更高的阈值，即以更高的标准评估受保护群体。这导致在情景2中，受保护群体中获批的人数减少，从而使真正率差距扩大。在第三种场景中，两家银行都对受保护群体不公平，导致真正率差距进一步扩大。在情景4中，一家企业对受保护群体不公平，而另一家企业对非受保护群体不公平，因此真正率差距缩小。

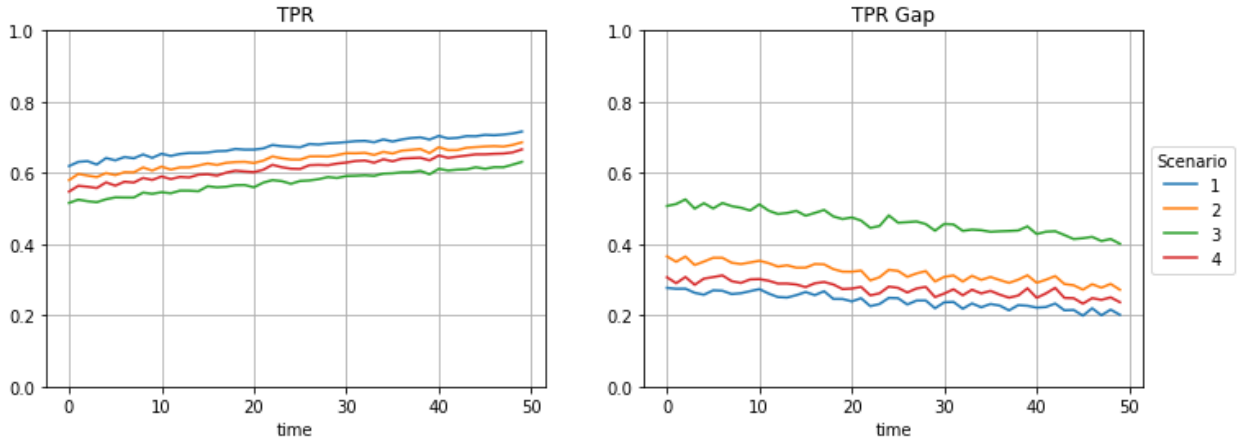


图5：企业表现出的显性偏差的影响。情景1中，两家企业均公平，受保护群体和非受保护群体的阈值均为0.8。情景2中，企业1不公平，对受保护群体的阈值较高，为0.9。情景3中，两家企业均不公平，对受保护群体的阈值较高，为0.9。情景4中，企业1对受保护群体的阈值较高，企业2对非受保护群体的阈值较高。

6 模拟研究 3

6.1 设置

为了模拟更真实的情景，本研究将企业数量从 2 家增加到 20 家，并将个体数量从 1000 人增加到 100,000 人。我们还放宽了所有个体都向所有企业申请贷款这一假设。通常，由于每次申请都会产生直接或间接成本，个体可能不愿意向所有企业提交申请。例如，货比三家申请贷款或许能获得更优惠的利率，但每次申请都会产生一项间接成本，即对申请人的信用报告进行硬查询，而这会降低申请人的信用评分。

6.2 结果

为了更好地评估现实情境中的公平性水平，我们首先测量基线情景（情景1）下的模型公平性和有效性，此时所有个体都会向全部20家企业申请。在这种情况下，我们观察到较高的真正例率（TPR）和较低的真正率差距。这是因为在这个规模较大的生态系统中，大多数人都能获得一家或另一家企业的批准。在情景2中，非受保护群体的人会向全部20家企业申请，而受保护群体的人只会向随机选出的3家企业申请。随着企业数量减少，批准率也随之下降，因此真正例率也降低。由于受保护群体的申请数量较少，他们受到的影响比非受保护群体更大，因此真正率差距也会扩大。在情景3中，我们模拟了这样一种情况：受保护群体选择的3家企业并非从生态系统中的全部20家企业里随机选出，而是从其中风险阈值较低的一半企业中选出。与情景2相比，真正例率和真正率差距均有所改善；这是因为选择阈值较低的企业会增加受保护群体获得批准的机会，从而提升真正例率并缩小真正率差距。在情景4中，我们还模拟了阈值较低的企业可能具有较高服务成本的情况，而这可能会增加出现负面结果的可能性。例如，在贷款业务中，风险阈值较低的银行可能会收取较高利率，进而提高违约概率。与情景3相比，我们观察到真正例率略有下降，真正率差距也有所扩大。从这些情景中可以推断，尽可能多地向企业申请有助于改善公平性，而政策制定者可以引入规则来消除申请成本障碍。如果没有此类规则，最好引导个体选择某一特定类型的企业，而非其他企业，即使其服务成本较高。

7 讨论

从自然语言处理模型到推荐系统，研究人员一直在研究算法，以检测和衡量其对受保护群体的偏见 [31, 32, 33]。这一系列研究对于增进我们对算法如何影响几乎所有领域中个体的决策过程的理解，具有不可估量的价值。受此启发，我们看到新一系列研究开始探讨开发抑制这些偏见的方法 [34, 25]。

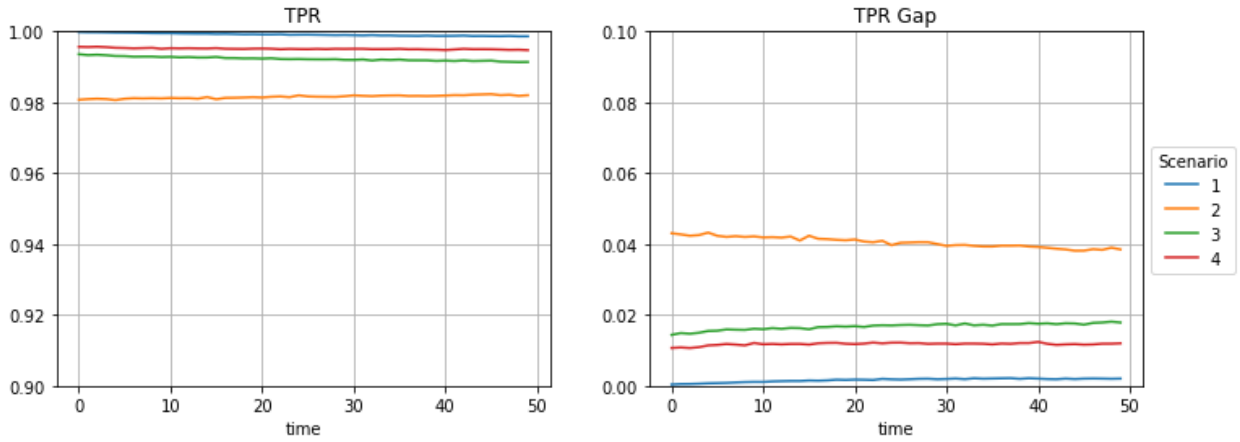


图6：个体选择性应用的影响。

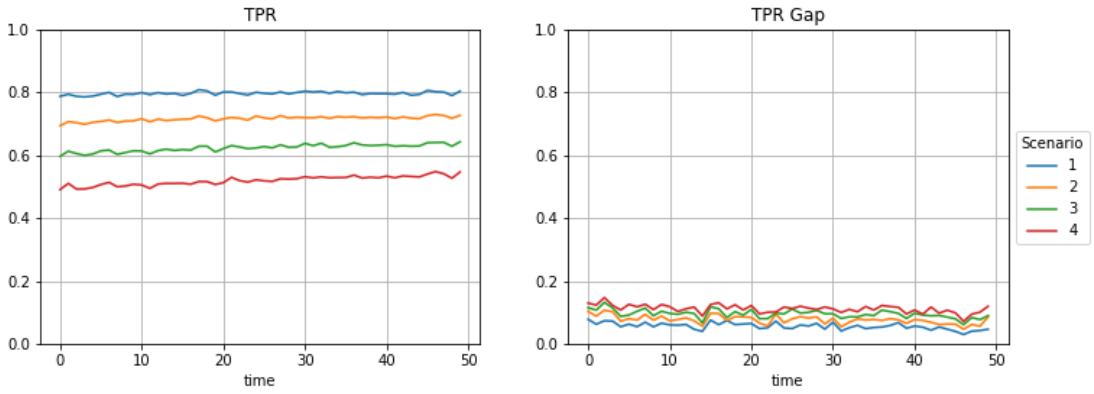
算法公平问题的严重程度也促使政策制定者制定立法，旨在限制此类算法在实践中的权力 [35]。目前，这通过强调人工裁量、减少决策自动化来实施。尽管这些解决方案有助于短期内解决问题，但这些方法可能会：1）增加实际使用和维护算法的相关成本；2）降低部署这些算法所带来的益处；3）将人为偏见引入决策过程。

鉴于这些问题在实践中的重要性，我们的模型和框架必须考虑到不同情景中发挥作用的整个生态系统。否则，看似良好的解决方案在实际部署时，可能无法实现我们所预期的公平结果。基于这一认识，我们通过开发和评估一个用于生态系统层面系统性公平的公平框架，拓展了公平性领域的相关文献，并主张在这一方向开展更多研究。此类场景能够提供宝贵的见解，为个体企业决策和政策举措提供指导。例如，在本文考察的简单场景中，我们通过系统性公平视角发现，特定企业较低水平的公平性可以通过市场中其他企业的设立得到弥补。

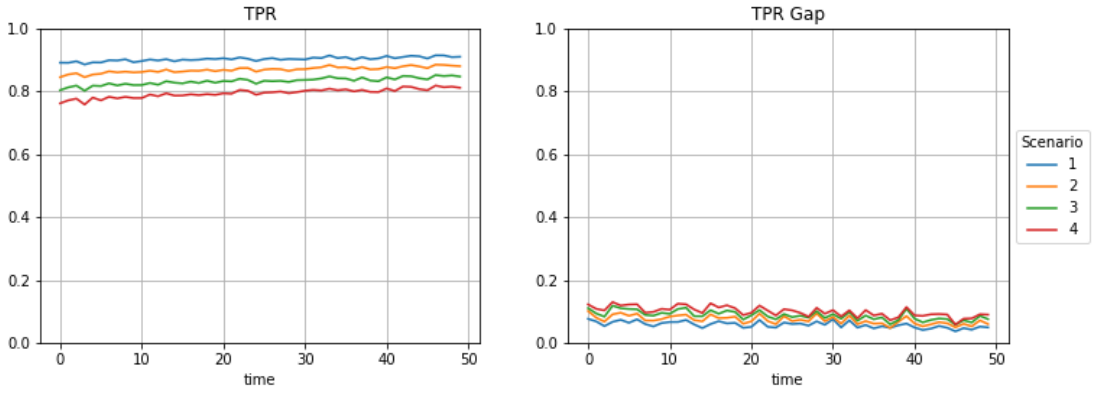
将公平性研究的视角拓展至系统层面，为多个研究方向打开了大门。作为本研究的后续工作，我们计划在生态系统中引入更多异质企业，并评估它们对企业层面和系统层面公平性的影响。我们还将在生态系统场景中，考察以往文献在单个企业层面提出的不同干预措施的有效性。虽然本研究成功地将借贷领域用作系统性公平框架的试验平台，但仍需更多研究来评估同一框架在其他领域中的适用性。

A 代表性不足的受保护群体

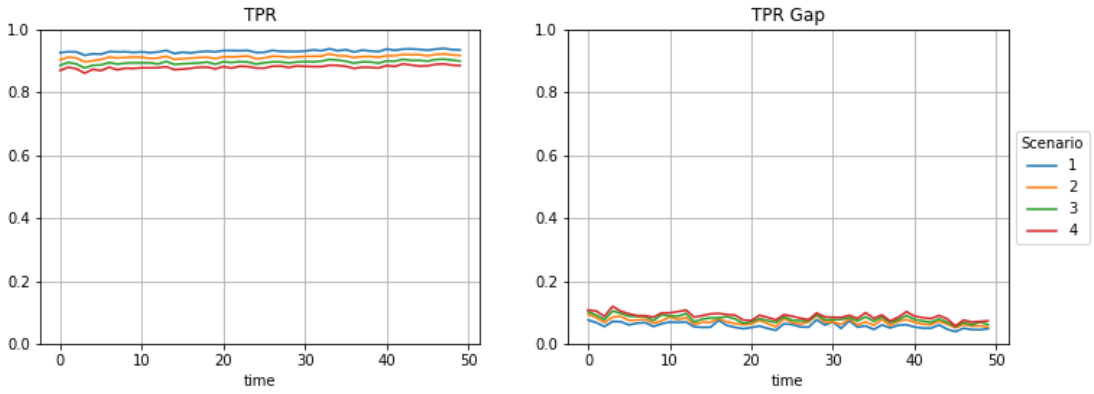
在论文中，我们假设受保护群体和非受保护群体在人口中的占比相同。尽管现实中，受保护群体通常占比不足。例如，美国人口普查局估计，非裔美国人占人口的 13.6%，西班牙裔或拉丁裔占 18.9% [36]。我们重新运行了模拟研究 1，将受保护群体设定为占人口的相同比例，并发现即使受保护群体代表性不足，我们的观察结果仍然成立。



(a) 单一企业、焦点企业 $S = 0.7$



(b) 多企业、焦点企业 $S = 0.7$、第二家企业 $\tau = 0.7$， $s = 0.7$

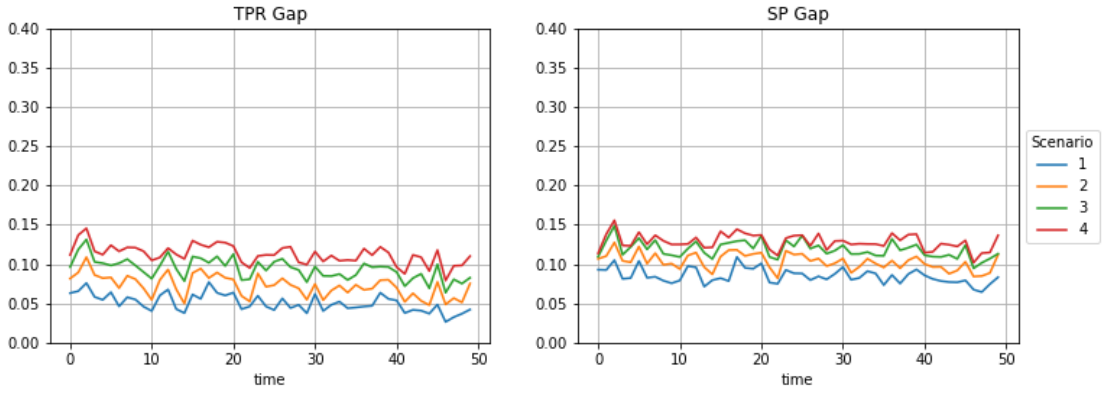


(c) 多企业、焦点企业 $S = 0.7$、第二家企业 $\tau = 0.6$， $s = 0.9$

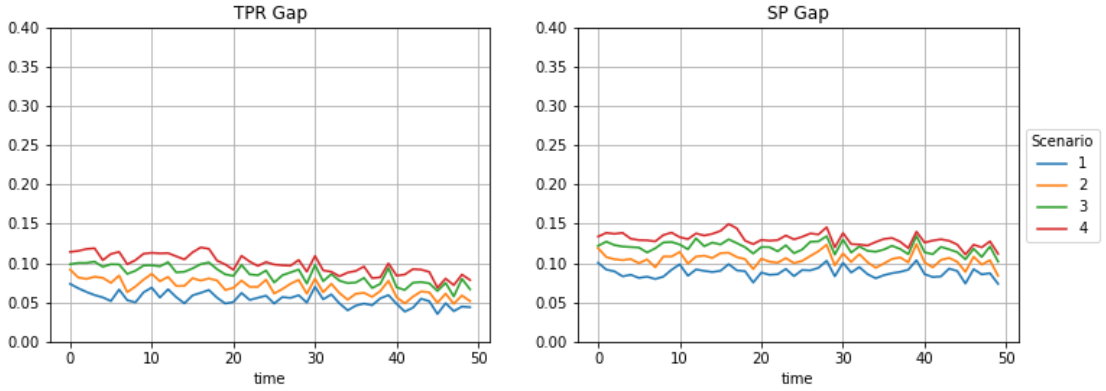
图 7：在焦点企业复杂度较低的单一企业和多企业场景中，受保护群体与非受保护群体的真正率及真正率差距的变化。所有面板中的情景 1、2、3 和 4 分别表示焦点企业的阈值为 0.5、0.6、0.7 和 0.8。

B 其他公平性度量下结果的可推广性

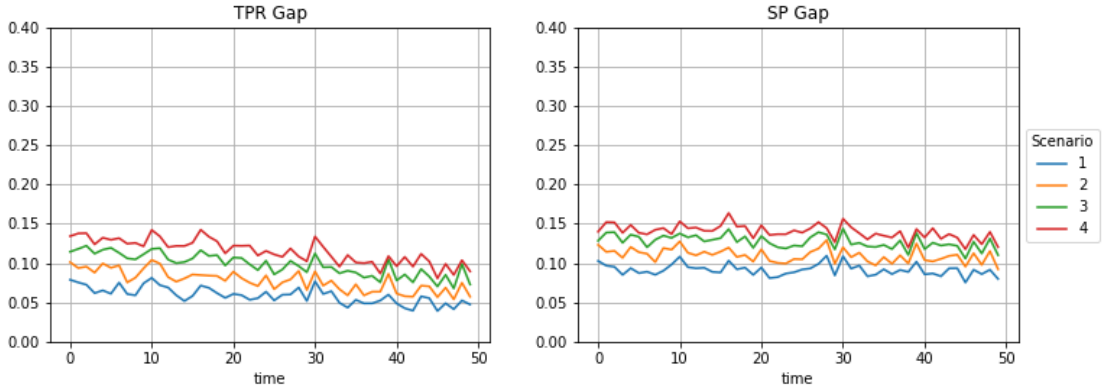
在我们的仿真研究 1 中，我们还测量了统计均等差距，将其作为另一项公平性度量。我们观察到，与真正率差距相关的讨论结果同样适用于统计均等差距度量。这一观察有助于证明，系统性公平现象能够推广到多种公平性度量，而不仅仅局限于机会均等度量。



(a) 单一企业、焦点企业 $S = 0.7$



(b) 多企业、焦点企业 $S = 0.7$、第二家企业 $\tau = 0.7$， $s = 0.7$



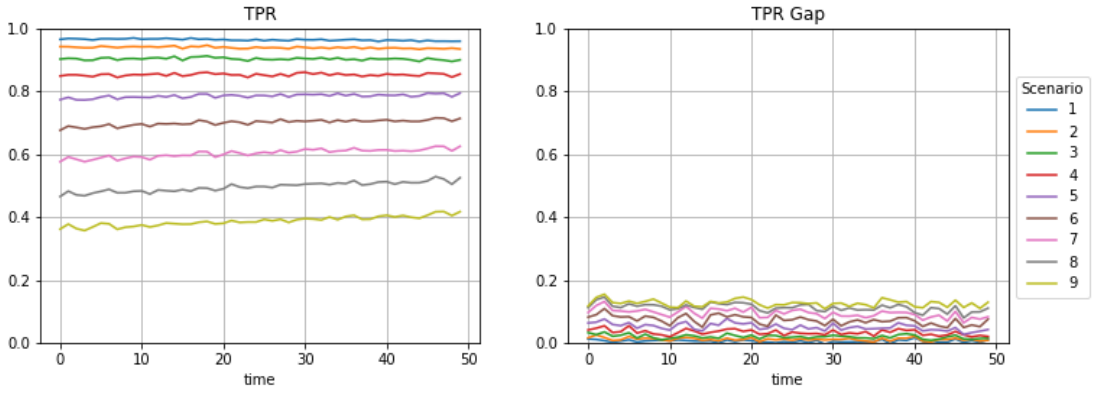
（c）多企业、焦点企业 $S = 0.7$、第二家企业 $\tau = 0.6$， $s = 0.9$

图8：在焦点企业复杂度较低的单一企业和多企业场景中，受保护群体与非受保护群体之间真正率差距和统计均等差距的变化。所有图中，情景1、2、3和4分别表示焦点企业的阈值为0.5、0.6、0.7和0.8。

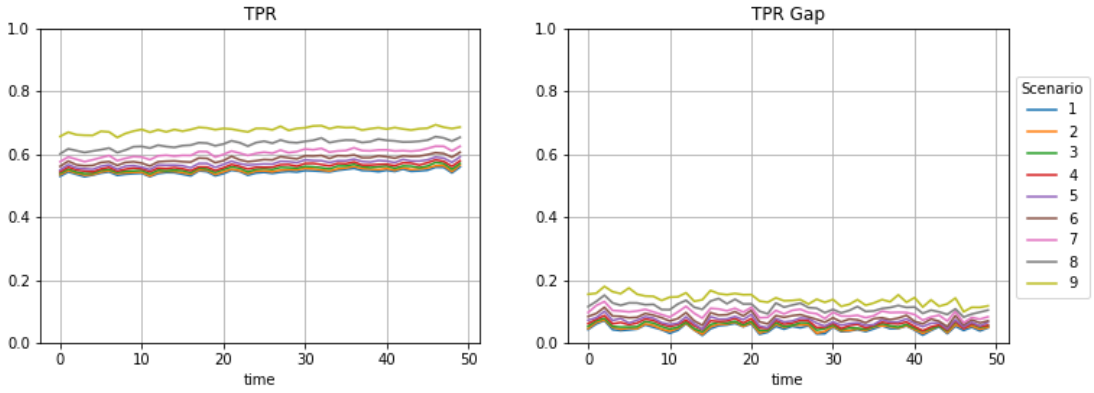
C 模型有效性与公平性：在更广泛的阈值和复杂度取值范围内

虽然复杂度和阈值参数都可以取 0 到 1 之间的任意值，但在模拟研究中，我们采用了适度的取值，通常介于 0.5 到 0.9 之间。对于银行而言，阈值为 0.5 意味着银行会批准估计还款概率为 0.5 的个人。基于这一点定义，我们认为银行的阈值低于这一数值的可能性极低。同样，复杂度为 0.5 意味着对个人质量的估计误差具有 0.5 的方差。对于质量这类取值范围在 0 到 1 之间的指标，算法的复杂度低于 0.5 的可能性同样极低。

然而，为了确保分析的全面性，我们针对覆盖整个可能范围的阈值和复杂度值，重复进行了模拟研究 1a，并观察到这一趋势在整个范围内持续存在。图 9a 表明，我们在模拟研究 1 中观察到的效应，即随着阈值降低，效率和公平性均得到提升，这一结论在整个阈值范围内均成立。图 **??** 表明，随着复杂度水平的提高，真正例率（TPR）也随之提高，但真正率差距也会扩大，因为银行能够以更高的准确度预测个人质量。



(a) 单一企业， $S = 0.7$在场景1中随 τ 从0.1变化至0.9的情况彩票开户



(b) 单一企业， $\tau = 0.7$，在场景1到9中， s 从0.1到0.9变化

图 9：在不同阈值和复杂度水平下，单一企业场景中受保护群体与非受保护群体之间真正例率（TPR）和真正率差距（TPR Gap）的变化。

参考文献

[1] Claire Cain Miller。算法能否比人类更好地招聘？——《纽约时报》。https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html，2015年。（访问日期：03/28/2022）。[2] Austin Waters 和 Risto Miikkulainen。GRADE：用于研究生招生的机器学习支持。《*AI Magazine*》，35(1):64–64，2014年。[3] Bart Baesens、Rudy Setiono、Christophe Mues 和 Jan Vanthienen。利用神经网络规则提取和决策表进行信用风险评估。《管理科学》，49(3):312–329，2003年。[4] Richard Berk。机器学习风险预测对假释委员会决策和再犯的影响评估。《实验犯罪学杂志》，13(2):193–216，2017年。[5] Julia Angwin、Jeff Larson、Surya Mattu 和 Lauren Kirchner。机器偏见——ProPublica。https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing，2016年5月。（访问日期：08/11/2021）。[6]Isobel Asher Hamilton。Sandra Wachter 表示，亚马逊的人工智能存在针对女性的偏见，这并不令人意外。https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10，2018年10月。（访问日期：08/11/2021）。[7] Robert Bartlett、Adair Morse、Richard Stanton 和 Nancy Wallace。金融科技时代的消费者贷款歧视。《金融经济学杂志》，2021年。[8] Cynthia Dwork、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Richard Zemel。认知公平。载于第三届理论计算机科学创新会议论文集，第214–226页，2012年。[9] Rich Zemel、Yu Wu、Kevin Swersky、Toni Pitassi 和 Cynthia Dwork。学习公平表征。载于国际机器学习会议，第325–333页。PMLR，2013年。[10] Tolga Bolukbasi、Kai-Wei Chang、James Y Zou、Venkatesh Saligrama 和 Adam T Kalai。男人之于计算机程序员，犹如女人之于家庭主妇？消除词嵌入中的偏见。神经信息处理系统进展，29，2016年。[11] Moritz Hardt、Eric Price 和 Nati Srebro。监督学习中的机会均等。神经信息处理系统进展，29，2016年。[12] Sam Corbett-Davies、Emma Pierson、Avi Feller、Sharad Goel 和 Aziz Huq。算法决策与公平成本。载于第*23*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第797–806页，2017年。[13] Matt J Kusner、Joshua Loftus、Chris Russell 和 Ricardo Silva。反事实公平。神经信息处理系统进展，30，2017年。[14] Joy Buolamwini 和 Timnit Gebru。性别阴影：商业性别分类中的交叉性准确率差异。载于公平、问责与透明度会议，第77–91页。PMLR，2018年。[15] Mehmet Eren Ahsen、Mehmet Ulvi Saygi Ayvaci 和 Srinivasan Raghunathan。当算法预测使用人工生成的数据时：一种用于乳腺癌诊断的偏差感知分类算法。《信息系统研究》，30(1):97–116，2019年。[16] Anja Lambrecht 和 Catherine Tucker。算法偏见？关于 STEM 职业广告展示中明显的性别歧视现象的实证研究。《管理科学》，65(7):2966–2981，2019年。[17] Gediminas Adomavicius 和 Mochen Yang。整合行为、经济和技术洞见，以理解和解决算法偏见：以人为本的视角。《*ACM*管理信息系统汇刊》（*TMIS*），2022年。[18] Till Speicher、Hoda Heidari、Nina Grgic-Hlaca、Krishna P Gummadi、Adish Singla、Adrian Weller 和 Muhammad Bilal Zafar。量化算法不公平性的一种统一方法：通过不平等指数衡量个体与群体不公平性。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第2239–2248页，2018年。[19] Alexander D’Amour、Hansa Srinivasan、James Atwood、Pallavi Baljekar、D Sculley 和 Yoni Halpern。公平性并非静态：通过模拟研究深入理解长期公平性。载于*2020*年公平、问责与透明度会议论文集，第525–534页，2020年。[20] Lydia T Liu、Sarah Dean、Esther Rolf、Max Simchowitz 和 Moritz Hardt。公平机器学习的延迟影响。载于国际机器学习会议，第3150–3158页。PMLR，2018年。

[21] Hoda Heidari、Vedant Nanda 和 Krishna P Gummadi。算法决策策略的长期影响：社会学习中的努力不公平与特征隔离。*arXiv*预印本 *arXiv:1903.01209*，2019年。[22] Sampath Kannan、Aaron Roth 和 Juba Ziani。平权行动的下游影响。载于公平、问责与透明度会议论文集，第240–248页，2019年。[23] Lydia T Liu、Ashia Wilson、Nika Haghtalab、Adam Tauman Kalai、Christian Borgs 和 Jennifer Chayes。当个体进行理性投资时，算法决策的差异化均衡。载于*2020*年公平、问责与透明度会议论文集，第381–391页，2020年。[24] Hussein Mouzannar、Mesrob I Ohannessian 和 Nathan Srebro。从公平决策到社会平等。载于公平、问责与透明度会议论文集，第359–368页，2019年。[25] Min Wen、Osbert Bastani 和 Ufuk Topcu。序贯决策中的公平算法。载于国际人工智能与统计会议，第1144–1152页。PMLR，2021年。[26] Lily Hu 和 Yiling Chen。劳动力市场长期公平性的短期干预。载于*2018*年万维网大会论文集，第1389–1398页，2018年。[27] Tatsunori Hashimoto、Megha Srivastava、Hongseok Namkoong 和 Percy Liang。在重复损失最小化中实现无人口统计特征的公平性。载于国际机器学习会议，第1929–1938页。PMLR，2018年。[28] Xueru Zhang、Mohammadmahdi Khaliligarekani、Cem Tekin 等。序贯决策中使用机器学习时的群体留存：用户动态与公平性的相互作用。神经信息处理系统进展，32，2019年。[29] Ninareh Mehrabi、Fred Morstatter、Nripsuta Saxena、Kristina Lerman 和 Aram Galstyan。机器学习中的偏见与公平性综述。*ACM*计算调查（*CSUR*），54(6):1–35，2021年。[30] Sahil Verma 和 Julia Rubin。公平性定义解析。载于*2018*年*IEEE/ACM*软件公平国际研讨会（*Fairware*），第1–7页。IEEE，2018年。[31] Ben Hutchinson、Vinodkumar Prabhakaran、Emily Denton、Kellie Webster、Yu Zhong 和 Stephen Denuyl。自然语言处理模型中的社会偏见对残障人士构成的障碍。*arXiv*预印本 *arXiv:2005.00813*，2020年。[32] Michael D Ekstrand、Anubrata Das、Robin Burke、Fernando Diaz 等。信息获取系统中的公平性。信息检索基础与趋势®，16(1-2):1–177，2022年。[33] Masoud Mansoury、Himan Abdollahpouri、Mykola Pechenizkiy、Bamshad Mobasher 和 Robin Burke。推荐系统中的反馈循环与偏见放大。载于第*29*届*ACM*信息与知识管理国际会议论文集，第2145–2148页，2020年。[34] Masoud Mansoury、Himan Abdollahpouri、Mykola Pechenizkiy、Bamshad Mobasher 和 Robin Burke。一种缓解推荐系统多方曝光偏差的基于图的方法。*ACM*信息系统汇刊（*TOIS*），40(2):1–31，2021年。[35] Ben Green。要求对政府算法进行人工监督的政策存在的缺陷。计算机法律与安全评论，45:105681，2022年。[36] 美国人口普查局。美国人口普查局 QuickFacts：美国。https://www.census.gov/quickfacts/ fact/table/US/PST045222，2022年7月。（访问日期：2023年4月11日）。