用户驱动的算法审计中的参与和劳动分工：普通用户如何协作揭示算法危害?

Rena Li*Sara Kingsley*
renal@andrew.cmu.edu
skingsle@cs.cmu.edu 卡内基梅隆大学 美国宾夕法尼亚州匹兹堡

Chelsea Fan†Proteeti Sinha†Nora Wai†卡内基梅隆大学 美国宾夕法尼亚州匹兹堡

Jaimie Lee‡
jaimiele@andrew.cmu.edu 卡内基梅隆大学 美国匹兹堡

Hong Shen
hongs@andrew.cmu.edu 卡内基梅隆大学 美国匹兹堡

Motahhare Eslami
meslami@andrew.cmu.edu
卡内基梅隆大学
匹兹堡，美国

Jason Hong
jasonh@cs.cmu.edu
卡内基梅隆大学
美国匹兹堡

摘要

近年来，出现了一种有趣的现象：用户聚集起来，审视自己日常生活中遇到的可能有害的算法行为。研究人员已开始从理论和实证层面理解这些用户驱动的审计，希望借助用户的力量检测有害的机器行为。然而，对于用户在这些审计中的参与及其劳动分工，我们知之甚少，而这对未来支持这些集体努力至关重要。通过收集并分析近期四起用户驱动审计案例中的 17,984 条推文，我们揭示了用户参与和投入的模式，尤其关注每起案例中的主要贡献者。我们还识别出用户生成的内容在这些审计中发挥的多种作用，包括提出假设、收集数据、扩大传播、情境化和推动升级。我们讨论了设计工具以支持用户驱动审计，以及支持那些为提高人们对算法偏见的认识而付出劳动的用户所带来的启示。

CCS 概念

• **以人为中心的计算**！**人机交互（HCI）**；人机交互中的实证研究。

关键词

算法审计；用户驱动的审计；Twitter；Twitter 图像裁剪；ImageNet Roulette；Portrait AI；Apple Card；性别；种族主义；偏见；用户驱动的审计；用户生成内容；内容创作者；内容劳动

ACM参考格式：
Rena Li、Sara Kingsley、Chelsea Fan、Proteeti Sinha、NoraWai、JaimieLee、Hong Shen、Motahhare Eslami 和 Jason Hong。2023年。《用户驱动的算法审计中的参与和劳动分工：普通用户如何协作揭示算法危害？》载于《*2023*年*CHI*计算机系统人因会议论文集（*CHI '23*），*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，纽约，美国，19页。https://doi.org/10.1145/3544548.3582074

1 引言

算法系统中偏见的存在催生了审计方法，通常由人工智能/机器学习专家主导，用于调查这些系统是否存在有害行为 [17, 52]。这些专家主导的审计技术已成功 fi发现并缓解了许多有害算法行为的案例；然而，这些技术仍面临若干限制。首先，专家并不总能意识到新出现的偏见正以新的方式影响边缘化群体。此外，专家用于调查偏见的技术，其设计目的并非必然是检测算法系 统在部署前或部署后的不可预测行为。相反，大多数由专家主导的审计方法都是为检测统计差异而开发的——例如，它们并非用于判断某个算法是否在图像中审查或有害地描绘边缘化群体，或在提供在线服务时是否如此。

近年来，一种有趣的现象逐渐出现：用户驱动的审计能够克服其中一些局限。在用户驱动的审计中，最终用户自发聚集起来，对算法系统开展审计，以揭示、审视并理解他们在日常生活中遇到的潜在有害机器行为 [44, 63, 65]。近期的例子包括：Twitter用户发现其图像裁剪算法存在种族偏见 [41],；小企业主聚集起来，调查Yelp可能对未在该平台投放广告的商家存在的偏见 [27],；内容创作者评估YouTube的算法为何取消发布在该平台上的LGBTQ内容的营利资格[44], ；以及用户测试谷歌翻译是否存在性别偏见 [56]。其中一些

*两位作者对本研究的贡献相同。
†作者对本研究的贡献相同。
‡作者负责数据可视化，并完成了重要的数据分析工作。

本作品采用知识共享署名-非商业性使用-禁止演绎 4.0 国际许可协议授权。

CHI '23，*2023*年*4*月*23*日至*28*日，德国汉堡
©2023年版权归所有者/作者所有。ACM ISBN
978-1-4503-9421-5/23/04。
https://doi.org/10.1145/3544548.3582074

用户驱动的审计促使大型科技公司改变或停止在生产环境中使用其算法 [10, 58]„ 或推动立法者提出新的公共政策，表明这些调查能够产生广泛的社会影响 [75]。

受这些发展的启发，研究人员开始从理论和实证两个层面探索用户驱动的审计，希望借助最终用户的力量来检测有害的算法行为（例如， [27, 28]）。具体而言，Shen and DeVos 等人 [65] 采用探索性案例研究，构建理论来解释用户如何通过日常与算法系统的互动，检测、理解并质询存在问题的机器行为，以及与专家主导的审计相比，用户驱动的审计有时为何能够取得更好的效果 ff。在另一项定性研究中，DeVos 等人（2022）发现，在用户驱动的审计过程中，用户会进行意义建构，例如，根据他们对算法系统的观察推演相关理论，以解释自己所感知到的偏见，并提出了此类意义建构过程的模式 [23]。

然而，尽管近期取得了这些进展，我们对这些算法审计中用户的参与模式和劳动分工仍知之甚少，尤其是在涉及相对大量用户的情况下。如果我们希望在未来帮助用户组织并支持他们开展此类算法审计，分析用户驱动的审计中的参与模式和劳动分工至关重要。因此，在本研究中，我们提出了以下三个研究问题：**(RQ1)**用户在用户驱动的算法审计中的参与模式是什么，**(RQ2)** 这些审计中的主要贡献者是谁，具体而言，他们在贡献数量以及向其关注者传播审计方面的表现如何，以及 **(RQ3)** 推文在这些审计中扮演了哪些不*ff*同的角色?

为回答研究问题，我们对 Twitter 上四个不同的用户驱动审计进行了比较分析。我们选择 Twitter 作为主要研究平台，因为它能让我们获取丰富的用户生成数据，从而轻松比较不同案例。此外，过去 Twitter 也一直是用户审计员分享对问题性机器行为担忧的主要社交媒体平台之一。例如，在 YouTube LGBTQ 内容取消盈利资格案例中，用户未能得到系统开发者的回应后，最终选择公开发布推文，讨论机器偏见。我们选取了四个备受关注的案例，每个案例的参与者都相对较多，且涵盖不同审计流程和起源，以便进行有效比较。这些案例包括前述的 Twitter 图像裁剪算法，该算法似乎在图像预览中表现出种族偏见；ImageNet Roulette 艺术项目让人们看到计算机视觉分类器对上传图像的识别结果，而这些结果往往会显示冒犯性标签；Portrait AI 的智能手机应用会运用不同的著名绘画风格重新绘制照片，并且常常将不同种族的人重新绘制成白人；以及 Apple Card，这款信用卡在已婚夫妇中往往给予男性远高于女性的信用额度。我们总共收集了 17,984 条相关推文，并对其中 1,800 条进行了定性编码和分析。随后，我们确定并分析了每个案例中的主要贡献者，并归纳出五类用于描述用户劳动分工的类别。最后，我们开发了一种机器学习分类器，用于标注其余推文，并分析用户参与审计活动的频率和持续时间。

我们的分析揭示了这四个案例中用户参与及参与度的模式。我们还确定了用户生成内容在 Twitter 上这些审计中发挥的五种主要作用，包括提出假设（推文推测观察到的机器行为为何发生）、证据收集（推文使用证据支持或反驳他人提出的假设）、扩散（帮助分享和传播他人所需信息的推文）、情境化（将正在进行的审计置于更广泛的社会、技术或文化语境中的推文），以及升级（主要针对审计中观察到的情况作出情绪化反应的推文）。我们的研究成果补充并拓展了以往文献对用户驱动审计的认识，表明：（1）与其着重于让用户深入、持续地参与审计，有时简单、简短的推文也能在表达愤怒和提高认识方面发挥关键作用，并可能促使算法运营者及其他主管部门（例如政府机构）推动系统层面的改变；（2）与其着重于提出假设和收集证据的科学方法，“扩散”和“升级”对于用户驱动的审计尤其关键，因为它们有助于提升更广泛的公众认知，并促使算法运营者和政府主管部门作出回应。

总而言之，我们的研究提供了三项主要贡献：

- 用户参与模式与互动的比较分析：我们提供了四个案例中用户参与总体模式的描述性统计，包括参与用户数量、用户参与审计相关活动的时长，以及每个案例中的主要贡献者。
- 用户驱动的审计中劳动分工的识别*fi*：用户生成内容我们识别出在全部四项用户驱动的审计中，用户推文发挥的五种一般性作用，并描述了各类作用内部的劳动细分，以及这些作用在不同案例中的差异。
- 设计启示：基于这些实证发现，我们提出了一系列设计建议，以便在未来更好地支持用户驱动的审计，并讨论与传统类型的算法审计相比，用户驱动的审计取得成功意味着什么。

2 相关工作

在本节中，我们概述两个领域的相关研究。首先，我们梳理现有的算法审计方法，并说明本研究在这一领域中的定位。接下来，我们回顾集体智慧和意义建构方面的现有研究，并阐述我们的用户驱动审计研究如何借鉴并推动这一研究方向的发展。

2.1 算法审计

过去几年间，各类机器学习算法因存在偏见和有害行为而受到广泛批评。对此，人机交互、计算机支持的协同工作和机器学习领域的研究人员开发了多种方法，用于检查和调查算法系统。这些方法通常称为“算法审计” [17, 52], , 指“反复查询某个算法，或依赖算法系统的软件应用，

并观察其输出，以便对该算法不透明的内部运作及其可能造成的外部影响得出结论”[52]。

以往研究已成功在多种算法应用领域中发现有害和歧视性行为，包括搜索引擎 [52, 54, 60], 在线广告 [45, 67], 面部识别 [10], 词嵌入 [7], 和电子商务 [40]。例如，Buolamwini 和 Gebru 对三种商业性别分类系统进行了审计，发现这些商业系统对深色皮肤女性的分类错误率高于白人 [10]。在另一项研究中，Sweeney [67] 审计了谷歌的搜索和广告系统，发现搜索与黑人相关的姓名时，显示的逮捕记录广告比搜索与白人相关的姓名时更多。这些研究大多由具备较高技术专长的专家研究人员开展并集中组织。

近年来，出现了另一种形式的算法审计：普通用户聚集起来，检测、推测并测试自己使用的系统中是否存在偏见 [27, 28, 65]。通过一系列探索性案例研究，Shen and DeVos 等人 [65] 提出了“日常算法审计”这一概念，用以描述用户通过日常与算法系统的互动，探查有问题的机器行为的过程。DeVos 等人 [23] 开展了一系列出声思维访谈、日记研究和研讨会，更深入地了解用户如何 fi发现并理解有害的机器行为。他们的 fi研究结果表明，用户的审计策略在很大程度上受到其个人经历以及接触社会偏见的影响。这些用户驱动的审计也促使专家更严格地测试嵌入软件应用中的底层算法。例如，受到用户识别 Twitter 裁剪算法中种族和性别偏见的 ff努力的启发，Yee 等人 [74] 提供了定量公平性分析，展示了 Twitter 模型对不同种族和性别子群体造成的差别影响程度。专家也开发了支持用户驱动审计的工具。例如，Lam 等人（2022）[49] 提出了一个系统，旨在“降低算法审计的高 ff努力门槛”并“为最终用户提供 ff支 [持，以便推进]审计过程”。

以上表明，专家主导与用户驱动是描述算法审计的一个维度。Costanza-Chock 等人 [17] 提出了另一个维度，即区分 fi第一方、第二方和第三方算法审计。第一方审计由创建算法的组织或开发者开展，明确目标是系统地评估算法是否存在潜在偏见和有害行为 [17]。第二方审计则由付费专家开展，他们能够访问算法和/或后端数据，完成审计后将 fi发现报告给 fi第一方，有时也会向公众报告。相比之下，第三方算法审计由与相关组织没有合同关系的独立实体发起，例如独立研究人员、记者或政府机构。

结合这两个维度可以看出，该领域过去的大部分工作都是由专家驱动的，涵盖 fi一方、第二方和第三方审计。本文将用户驱动的审计视为一种新型的第三方算法审计。不过，一个主要区别在于，这类用户驱动的第三方审计采用的形式远不如专家驱动的审计有结构。另一个显著区别是，参与第三方用户驱动审计的众多用户可能没有明确一致的目标。尽管如此，我们考察的四个案例中，至少有三个成功促成了某种改变或干预。我们还注意到，也可以开展用户驱动的 fi一方审计，例如公司成员可能会在内部试用某个系统 [17,65], ；也可以开展用户驱动的第二方审计，即公司雇用众包工作者系统地测试某个系统（参见[11]，这是一个可助力此类工作的工具示例）。不过，这两类审计不在本文的讨论范围之内。

尽管近期取得了这些进展，目前仍缺乏对用户驱动审计在现实世界中的诸多动态特征进行的实证分析。以往研究 [65] 表明，用户驱动审计有不同类型，涵盖不同程度的算法专业知识、集体性和自发性。然而，我们目前对用户如何参与不同类型的算法审计了解有限。例如，典型的参与模式是什么？用户会贡献哪些类型的用户生成内容？理解这些问题有助于未来分析用户驱动审计，也有助于开发支持这类审计的工具。本文采用混合方法，通过实证比较用户在 Twitter 上集体开展的四种不同审计案例，为这一新兴研究方向增添了成果。结果表明，尽管这些案例在多个方面各不相同，但在参与者和贡献类别方面也存在许多相似之处。

2.2 集体智慧与意义建构

Weick [71] 将意义建构定义为“将事物置于框架之中、加以理解、消解惊讶、建构意义、通过互动寻求相互理解并形成模式”。这一概念被广泛用于考察一群人如何协同工作、彼此借鉴发现与见解，并朝着共同目标努力——有时能比各自独立工作的领域专家取得更好的成果。以往研究探讨了多种情境中的集体智慧与意义建构，包括图像标注、 [70], 知识图谱绘制与策展（例如在 Wikipedia 上） [30, 33, 38, 39], 科学研究与协作 [16, 18], 以及社交商务 [13]。另一类与意义建构相关的文献探讨用户如何形成“民间理论”，以解释他们使用计算机和算法系统时的体验与观察 [21, 37, 43]。以往研究考察了这些民间理论和既有认知如何塑造并影响人们对算法系统的态度与行为 [6, 9, 22, 25, 36, 68, 72]。越来越多的此类研究聚焦于用户关于算法偏见的理论 [22, 43, 44]。近期有关用户驱动算法审计的研究也讨论了民间理论在用户提出假设和测试算法系统潜在偏见时所发挥的作用 [65]。然而，在这一意义建构过程中，用户在评估和检测潜在算法偏见时承担何种角色，仍鲜为人知。

我们的研究旨在通过考察 Twitter 上四个不同的算法审计案例中的用户参与和劳动分工，补充这一研究脉络。具体而言，我们希望了解 Twitter 用户如何协作提出假设，

检验并理解社交媒体平台上存在问题的算法行为，同时提高人们对有害偏见问题的认识。

3 背景

此处，我们介绍所研究的 Twitter 上四个审计算例的背景信息，分别是 Twitter 图像裁剪、ImageNet Roulette、Portrait AI 和 Apple Card。选择这四个案例有多方面的原因 [22, 57]。我们选择 Twitter 作为主要研究平台，因为它让我们能够获取丰富的用户生成数据，并便于比较不同案例。我们选择了可见度相对较高、且有相当数量的人讨论算法偏见问题的案例，因此能够获得较大量的用户生成数据进行分析。我们还寻找涵盖不同起源和审计流程的案例。例如，Twitter 图像裁剪、Portrait AI 和 Apple Card 案例主要是自下而上的：最初并不清楚算法存在偏见，用户自发聚集起来测试偏见。相比之下，ImageNet Roulette 更偏向自上而下，其网站创建者设计了这一工具，旨在揭示算法中潜藏的偏见。此外，Twitter 图像裁剪案例持续时间较短，聚焦于某个特定问题；而 Portrait AI 则长期存在，没有特定的关注重点。另外，Twitter 图像裁剪、ImageNet Roulette 和 Portrait AI 主要关注图像，而 Apple Card 关注的则是用户体验和证言。

根据该领域以往的研究 [23, 65],，我们将这些案例视为用户驱动的算法审计，因为它们fi首先是由算法系统的用户开展的。用户驱动的审计可能具有不同程度的算法专业知识、集体性和有机性；然而，它们与专家主导的审计有所不同，因为所报告的有害机器行为只有在真实且具体的人工智能使用情境中才能被检测到 [65]。我们还沿用 [65],所提出的生命周期，考虑到用户驱动的审计所涵盖的一系列活动，从较高层面来看，包括以下阶段：启动、提高公众意识、提出假设与测试，以及补救。然而在实践中，用户驱动的审计可能遵循非线性路径；有些审计可能跳过某些阶段，有些则可能在完成所有阶段之前结束，正如我们对这四个案例的分析所示。

3.1 Twitter 图像裁剪案例

为了让网站上发布的照片缩略图尺寸保持统一，Twitter 过去使用图像裁剪算法自动裁剪用户的照片。2020年9月18日，博士生 Colin Madland 发现，他发布的一张自己与一位黑人同事的合照似乎总是优先显示他 [Colin 的] 脸（即使 fi换图像的方向和顺序也是如此），这表明该算法偏向肤色较浅的面孔。[1] 这条 Twitter 帖子引发了轩然大波，促使许多用户自行测试，并 fi发现 Twitter 图像裁剪算法存在偏见的其他案例。一项测试是上传一张同时包含巴拉克·奥巴马总统（深色皮肤）和参议员 Mitch McConnell（浅色皮肤）的照片，采用各种布局，并在两张脸之间留出大量空白区域，以迫使算法做出选择单张脸。有些图像将 Obama 放在上方、麦康奈尔放在下方，另一些则反过来。裁剪时，无论fi配置如何，算法似乎都偏向麦康奈尔。这种fi特定测试方法，即在同一张图像中仅放置两张相距很远的面孔，很快就被其他审计员 [41]采用。

在 10 月 1, 2020 [3], , Twitter 在一篇博客文章中披露，其使用的是显著性算法，也就是说，该算法会聚焦于人们最可能注意到的图片局部 [3]。Twitter 承认，这种算法可能造成伤害，包括基于人口统计差异的不平等待遇、物化偏见，以及用户无法在没有算法替自己做决定的情况下自由表达。针对公众对有偏见的图像裁剪算法的强烈抗议，Twitter 作出了公开道歉。

总体而言，这项审计受到了大众媒体的广泛报道（例如，[2, 41, 42]）。很大程度上由于这项审计结果的公开传播，Twitter 于 2021 年 5 月彻底改造了图像裁剪功能，移除了算法，并允许用户上传任意标准宽高比的图片。[2] 推文预览还会向用户展示图片发布后的效果。此外，Twitter 宣布将在 2021 年 DEFCON AI Village 举办软件行业首个算法偏见漏洞赏金竞赛 [15], 。Twitter 将其图像裁剪代码开源，并邀请参与者提交偏见评估，以帮助发现更广泛的问题。2021 年 9 月，Twitter 的研究人员分享了竞赛结果，并承认即使进行专业审计，也不可能预见所有潜在问题，从而凸显了直接听取使用其产品的社群反馈的重要性 [48]。

3.2 ImageNet Roulette 案例

ImageNet Roulette [19] 是一项艺术展览，于2019年9月中旬公开推出，此前实体展览已于3月开幕。该展览旨在揭示使用 ImageNet数据集 [20]的计算机视觉算法中存在的潜在偏见。使用 ImageNet Roulette时，用户可以将自己、朋友甚至公众人物的照片上传到项目网站 fi。该工具会返回一张经过修fi改的图片，标出检测到的人脸，并显示描述照片中人物的标签。项目网站还设有一项功能，允许用户在社交媒体上分享结果，包括Twitter。[62]一些用户认为返回的标签带有种族歧视和性别歧视。Twitter上分享了其中一些分类fi结果，以及用户对所见内容的反应。

公开推出约两周后，展览创作者Trevor Paglen和Kate Crawford认为，这项实验已经达到了目的，于是于2019年9月27日关闭了项目网站。受该项目及其引发的关注影响，ImageNet中约有60万张图片被移除 [62]。

3.3 Portrait AI 案例

Portrait AI 是一个网站和智能手机应用，其最初版本使用人工智能技术修改上传的图像，使其看起来像一幅18世纪欧洲绘画。[3] Android 和 iOS 智能手机应用于2020年2月前后发布；截至本文撰写时，网站和应用

[1] https://twitter.com/colinmadland/status/1307111816250748933

[2] https://twitter.com/Twitter/status/1390026628957417473

[3] https://portraitai.app/

仍在运行，并已陆续增加许多其他艺术风格和视觉效果。Portrait AI 似乎由一位开发者独自创建，而 Android 应用的安装次数已超过一百万。

2020年5月左右，一些用户开始在 Twitter 上表达对 Portrait AI 的担忧，指出它会把有色人种转变成白人。[4] 截至本文撰写时，这些问题仍未得到解决，用户仍会偶尔在 Twitter 上发布他们的观察结果。不过，iOS 和 Android 版本的应用描述中都新增了一条说明："致有色人种用户：我们非常抱歉，我们的人工智能主要是用欧洲族裔人士的肖像进行训练的。我们计划很快 fix这一问题。"[5]

与上文概述的前两个案例不同，Portrait AI 是一个现有系统，许多用户都曾偶然发现其中可能存在偏见，但这些审计尚未促成多少新闻报道，也未推动算法发生改变。大多数 PortraitAI 用户仍在继续使用该系统并发布生成结果，却很少提及其中带有偏见的行为。

3.4 Apple Card 案例

苹果公司和高盛于 20, 2019 [1] 发布了新信用卡 Apple Card。几个月后，在 2019年11月7日，David Heinemeier Hansson 分享了自己和妻子使用 Apple Card 的糟糕经历，称尽管他们"fi联合报税申报、居住在夫妻共同财产州，而且结婚多年"，他的信用额度仍是妻子的 20 倍。他认为该程序存在性别歧视；当被告知无人获准透露评估流程时，他感到十分沮丧，对方还称这"只是算法"。[6] Hansson 发布的推文串引发轩然大波，其他消费者纷纷讲述类似经历，其中包括苹果公司联合创始人 Steve Wozniak。[7] 后来，Apple Card 人工提高了他们的信用额度[8]。

Apple Card 的发卡银行高盛回应称，其流程不存在偏见，因为在申请流程中，他们不会收集或考虑性别或婚姻状况 [47]。后来，纽约州金融服务局于 2019年11月9日 对公众的强烈反响作出回应，并表示将展开自己的调查。[9] 大约一年半后，纽约州金融服务局认定并未对女性实施非法歧视，相关问题源于产品发布管理不善 [55]。尽管有这些 fi调查结果，Twitter 用户仍不断报告称，自己与他人的 Apple Card 申请结果和信用额度存在差异，并表示他们认为高盛的算法表现出性别偏见。TechCrunch 记者 Liz O'Sullivan 进一步指出：

> "我完全确信，高盛/苹果公司的算法存在歧视，市场上如今的每一种信用评分和承保算法也都如此。我同样毫不怀疑，如果研究人员获准访问验证这一主张所需的模型和数据，这些算法就会原形毕露。我之所以知道这一点，是因为纽约州金融服务局部分公开了审查高盛算法的方法论；正如你所料，他们的审计远未达到当今现代算法审计人员所遵循的标准。"[10]与 Portrait AI 类似，Apple Card 是一款底层算法被用户指称存在偏见的产品。与 Twitter 图像裁剪案例类似，用户成功引起了足够关注，促使有关机构（纽约州金融服务局）展开调查。然而，与上述所有案例不同，Apple Card 审计是由一些颇具影响力的人物引发的，他们在 Twitter 上拥有大量受众，并带来了大量媒体报道和公众关注。这些有影响力的人物包括审计的发起者、Ruby on Rails 的创作者 David Hansson；著名音乐人 John Legend；Steve Wozniak；以及 Apple Card 本身。此外，与其他案例中大量的图像不同，用户审计所提供的大部分证据由用户体验报告和证言构成。

4 方法论

本文收集并分析了 Twitter 数据，因为参与用户驱动审计的人们曾利用该平台分享信息、提高公众意识，并要求平台运营方和相关部门回应偏见问题。即使算法偏见发生在其他平台上，用户也主要通过 Twitter（而非发生偏见的平台）来实现这些目的 [44]。同样，参与本文所分析的用户驱动审计案例的用户，也在 Twitter 上分享有关算法偏见的信息。本文涉及的三个案例中，尽管核心算法偏见发生在其他平台上，但我们没有收集这些平台的数据，因为这些平台不具备信息分享或社交媒体功能，无法让用户彼此分享有关偏见的信息，也无法让此前未使用过这些算法的人了解相关情况。因此，下文将详细说明我们如何收集并分析关于四起算法偏见案例的 Twitter 数据。

方法论概要：在详细说明所采取的程序之前，我们先在此对方法论作一概述。为了收集与本文介绍的四个用户驱动审计案例相关的 Twitter 数据，我们采用了以下方法。首先，我们通过搜索相关关键词从平台收集推文，然后递归收集原始推文中引用的其他推文。由此得到初始数据集，共包含 33,597 条推文。随后，我们构建了一个相关性分类器，筛除不相关的推文，最终得到包含 17,984 条推文的数据集。接着，我们分析了这些数据，考察最具影响力的发推者，以及不同推文如何在发现、讨论和揭示可能有害的算法行为方面发挥不同作用。为此，我们对整个语料库中的一个子集进行了人工编码，通过迭代方式分析数据并逐步制定和完善编码方案。最后，我们构建了一个机器学习分类器，将编码应用于整个语料库。出于隐私考虑，我们按照 Fiesler 和 Proferes 的建议，删去了本文所展示推文中的用户名和网址 [29]。例外是背景部分（第 3 节），因为其中的原始推文及其发布者姓名已被大众媒体广泛报道。

[4]例如，参见 https://twitter.com/blutmut/status/1263757322633052162

[5] https://apps.apple.com/us/app/portraitai-classic-portrait/id1474884190

[6] https://twitter.com/dhh/status/1192540900393705474

[7] https://twitter.com/stevewoz/status/1193424787248279552

[8] https://twitter.com/dhh/status/1192944667202998272

[9] https://twitter.com/LindaLacewell/status/1193183785581498369

[10]https://techcrunch.com/2021/08/14/how-the-law-got-it-wrong-with-apple-card/

下面，我们将详细介绍这些步骤。

4.1 推文数据收集

在本节中，我们将详细介绍数据收集方法。对于所研究的每个用户驱动审计案例，我们都采用以下方法从 Twitter 收集数据。首先，我们使用 Python 工具 snscrape11 手动搜索 Twitter；该工具通过 Twitter 的 REST API 请求数据。snscrape 工具以关键词和时间范围为输入，搜索包含相关短语和话题标签的推文，然后返回一组相关推文的 URL。我们根据对这些案例的了解，并通过查阅有关这些案例的大众媒体文章（例如“Twitter Cropping”和“Image Cropping”），反复完善用于搜索的关键词、短语和话题标签列表。

为确定每个案例的时间范围，我们查阅大众媒体和其他来源，以大致确定开始日期和结束日期；随后，我们使用这些开始和结束日期来调整输入 snscrape 的日期。对于每个用户驱动审计案例，我们都特意选择了稍宽的时间范围，以更好地确保尽可能全面地收集数据集

由于 *snscrape* 只返回 URL，我们编写了一个自定义 Python 脚本来收集每条推文的其他信息，尤其是推文内容、对话 ID，以及“点赞”和“转推”的数量。我们还编写了另一个独立的自定义 Python 脚本，递归搜索并识别初始收集的推文集中所引用的任何推文的推文 ID。完成这一步后，我们共获得 33,597 条推文。请注意，Twitter 裁剪案例是最先收集的数据集，由另一位研究人员按照与上述略有不同的流程和脚本收集推文数据和互动数据。不过，我们在另外三个案例中收集了相同类型的数据，并认为流程或所收集的数据不存在重大差异。

4.1.1 研究软件的公开获取*ft*。 我们用于收集、解析案例相关 Twitter 数据并将其整理为数据框的 *snscrape*脚本和 Python 笔记本副本已发布于：https: //userdrivenaudits.github.io

4.2 用于筛除不相关推文的相关性分类器

本节详细介绍了我们为判定推文是否与用户主导的审计案例相关而采取的分类流程。初步检查抓取到的数据时，我们发现有相当多的推文并未讨论或包含与四个审计案例中任何一个相关的内容。例如，在 ImageNet Roulette 案例中，一条不相关的推文写道：“简直太迷恋这个 [已删去网址]了。”另一条来自 Apple Card 案例的不相关推文写道：“为什么那条狗和澳大利亚原住民一起走，他却在把他们往后排，带着偏见，说原住民购物卡上写着一个苹果不好，所有苹果都不好。”为了解决这个问题，我们开发了一个相关性分类器，以筛除可能的噪声。

为了构建这一相关性分类器，我们首先从原始的 33,597 条推文中抽取了 1,200 条作为样本。该样本包含来自四个用户驱动审计案例的推文。在这些样本中

[11]https://github.com/JustAnotherArchivist/snscrape

推文中，四名编码员通过协作方式对500条推文进行编码，将其标记为“相关”或“不相关”，同时讨论并反复完善相关性的定义。完成这500条样本的编码后，计算了编码员间信度（0.848），以确保所有编码员在编码其余样本之前达成基本一致。在确认一致性高于0.70 [50]后，剩余的700条推文由四名编码员分别独立编码。

为 1,200 条推文完成编码后，我们使用这些标注数据训练分类器，帮助我们筛除整个数据集中的噪声。为了构建这个相关性分类器，我们将人工编码的数据集划分为训练集和测试集，其中 70% 的数据集（840 个实例）用于训练，30% 的数据集（360 个实例）用于测试。我们进行了实验，并评估所测试模型的性能，以选出用于构建相关性分类器的模型，即支持向量机（SVM）模型。我们为模型添加并使用了额外特征：一元词组、二元词组、三元词组和行长度。最后，我们使用 10 折交叉验证对最终模型进行训练、测试和运行。相关性分类器最终的准确率为 88%，卡帕值为 0.77。使用这个最终的相关性分类器，我们筛除了所有不相关推文，最终得到 17,984 条相关推文。在 4.5 节中，我们提供了用于构建相关性分类器的开源软件链接，并报告模型中采用的特征工程 步骤和参数，以便他人复现我们的工作。

4.3 主要贡献者的数据分析

我们还想调查四个案例中各自的主要贡献者是谁。为了解用户通过哪些方式参与审计，我们将其分为两类并进行排名：(a) 内容生产者，即在某个案例中发布推文数量最多的用户；(b) 内容传播者，即在某个案例中其推文获得的点赞和转推参与度最高的用户。对于内容生产者，我们以推文数量作为衡量其为用户驱动的审计投入劳动量的替代指标。同样，对于内容传播者，我们旨在了解谁帮助将案例传播给更广泛的用户群体。在分析中，我们按参与者 ID 对收集到的推文数据进行分组，以汇总每位用户发布的所有推文；随后按推文总数（内容生产者）或点赞数与转推数之和（内容传播者）降序排列用户，并分别选出排名前十的用户。将主要贡献者归类为内容生产者和内容传播者后，我们进一步查看这些高频发推用户的 Twitter 个人资料，以分析其可信度和影响力。如果用户展示出相关研究、科学、技术、工程和数学或技术背景 [65]，我们就将该账号标记为专家 。如果该账号拥有 1,000+ 关注者，我们也将其标记为影*fl*响者/名人 。在这一分类体系中，同一用户既可能是专家，也可能是影响者/名人。

我们还将每个用户账户标记为个人、新闻、公司、机器人或其他。我们通过考量账户简介并大致了解帖子模式来作出这些判断。例如，在 Apple Card 案例中，我们标记为机器人的头号内容制作人，其账户简介如下：

“一个汇总美国保守派新闻媒体头条的 Twitter 机器人。”这是我们在数据集中标记的唯一一个机器人（还有一位作者使用 https://botometer.osome.iu.edu 检查主要贡献者是否为机器人）。

4.4 推文数据分析

为了分析用户驱动审计的动态，我们遵循 DeVito 等人 [22] 的做法，对数据集进行了两阶段编码。首先，我们随机选取了 1800 条推文，由人工编码员据此制定编码方案。其次，我们使用这 1800 条推文训练了一个机器学习分类器 fi，并用它对整个数据集进行分类。

4.4.1 编码流程。 在 fi第一阶段的人工编码中，我们制定了一套互斥的编码方案，围绕用户在算法审计中承担的不同类型的劳动，以及如何将这些劳动划分为不同角色或劳动分工展开（见表 3）。我们所说的劳动分工，是指用户的内容在审计中实现的任务，例如“提出假设”“数据收集”和“升级”。整套方案由两位主要作者和两位本科生研究助理通过反复讨论制定，并在数次大型团队会议上由研究团队审阅。

编码方案通过演绎与归纳思维共同制定。我们的编码方案还受到 [65], 的启发，该文将算法审计的生命周期描述为启动、提高意识、提出假设与测试以及补救。我们 fi 首先对 fi 前 900 条推文应用主题分析 [8] ，将推文分成小批次进行编码，并每周召开会议，复核编码员难以分类的推文；每次讨论后，我们都会完善代码本。所有编码员完成 400 条推文的编码后，我们举行了一次深入讨论，复核一致性较低的推文，并据此确定了最终的编码方案 fi。完成 900 条推文的编码后，四名编码员中的两名继续编码另外 900 条推文，最终形成一个包含 1800 条推文的数据集，用于训练机器学习分类器fi。

4.4.2 分类流程。 为了从更大范围了解用户在全部四个案例中的参与情况，接下来，我们构建了一个机器学习分类器fi，使用相同的编码方案对整个数据集进行分类。将 1800 条推文的样本（见上文）分为训练集和测试集，其中 70% 的数据集（N=1,260）用于训练，30% 的数据集（N=540）用于测试。我们采用 10 折交叉验证，使用朴素贝叶斯模型、支持向量机 (SVM) 和序列最小优化 (SMO) 算法开展实验，最终选择了性能更优的支持向量机分类器fi。该支持向量机分类器fi还使用了一些额外特征：一元词组、二元词组、三元词组、行长度、可伸缩模式（带间隔的 n 元词组），并纳入标点符号以保留话题标签。最终的分类器fifi准确率为 88%，卡帕值为 0.849，并用于对所有相关数据集进行分类。

4.5 用于构建分类器的数据集、参数和软件

训练、测试和最终数据集。我们使用不同的训练和测试数据集来构建相关性和劳动分工分类器，并生成用于所呈现分析的最终数据集本文中每个分类器的训练数据集和测试数据集，以及我们的最终数据集，均已发布在网上：https://userdrivenaudits.github.io。我们遵循了基于 Casey Fiesler 的研究成果制定的 Twitter 数据伦理使用指南，包括删除用户名和 URL。

软件： 我们使用开源软件 LightSide [12] 训练并测试了用于构建相关性与劳动分工分类器的模型性能。我们在本项目的网站 13 上公布了训练模型和构建这两个分类器所用的参数。任何人都可以下载并使用 LightSide 和我们的研究数据，复现本研究采用的流程。

5 发现

本节将介绍我们的分析结果。首先，我们呈现四个案例的描述性统计，重点关注用户参与和参与度的模式。其次，我们分析四个案例中的主要贡献者，重点考察主要内容贡献者（按贡献数量计算）和主要内容传播者（按向其关注者传播审计的程度计算）。最后，我们分析用户推文在这些审计中发挥的五种主要作用，并描述各作用内部的劳动分工细分，以及这些作用在四个用户驱动的审计案例之间的差异。

5.1 审计案例中的参与模式与参与度（RQ1）

表1呈现了四个用户驱动审计案例中用户参与和参与度的描述性统计。总体而言，我们发现：（1）四个案例中有三个（Portrait AI 除外）呈现出相似的用户参与模式：推文活动最初出现大幅爆发，随后进入低活跃度的长尾阶段；（2）在全部四个案例中，绝大多数参与者仅贡献了一条推文。

在四个案例中，Twitter 裁剪案例中用户贡献的推文总数最多（11,323 条），其次是 ImageNet Roulette（2,813 条）、Apple Card（2,408 条）和 Portrait AI（1,440 条）。其次，用户每日发布推文数量最多的时间点，即参与度“峰值”日，在 Twitter 裁剪案例中出现在一天之内，在 Apple Card 中出现在一周之内，在 ImageNet Roulette 中约为 6 个月后，而在 Portrait AI 中则超过一年。需要注意的是，对于 ImageNet Roulette，我们发现的第一条相关推文发布于其公开发布日前约 6 个月，当时它仍是一件实体展览。因此，ImageNet Roulette 的峰值实际上代表的是其公开发布日。

接下来，我们发现四个案例中的三个——Twitter 图像裁剪、ImageNet Roulette 和 Apple Card——呈现出相似的用户参与模式：推文数量一开始激增，随后用户活动进入长尾低迷期（见图1）。一种可能的解释是，这些审计分别引起了名人、公众人物和其他影响者的广泛关注，进而促使机构或技术平台运营者采取干预措施，解决用户提出的担忧。例如，在 Twitter 图像裁剪案例中，用户参与度的峰值出现在一名用户

[12]http://ankara.lti.cs.cmu.edu/side/download.html [13] https://userdrivenaudits.github.io

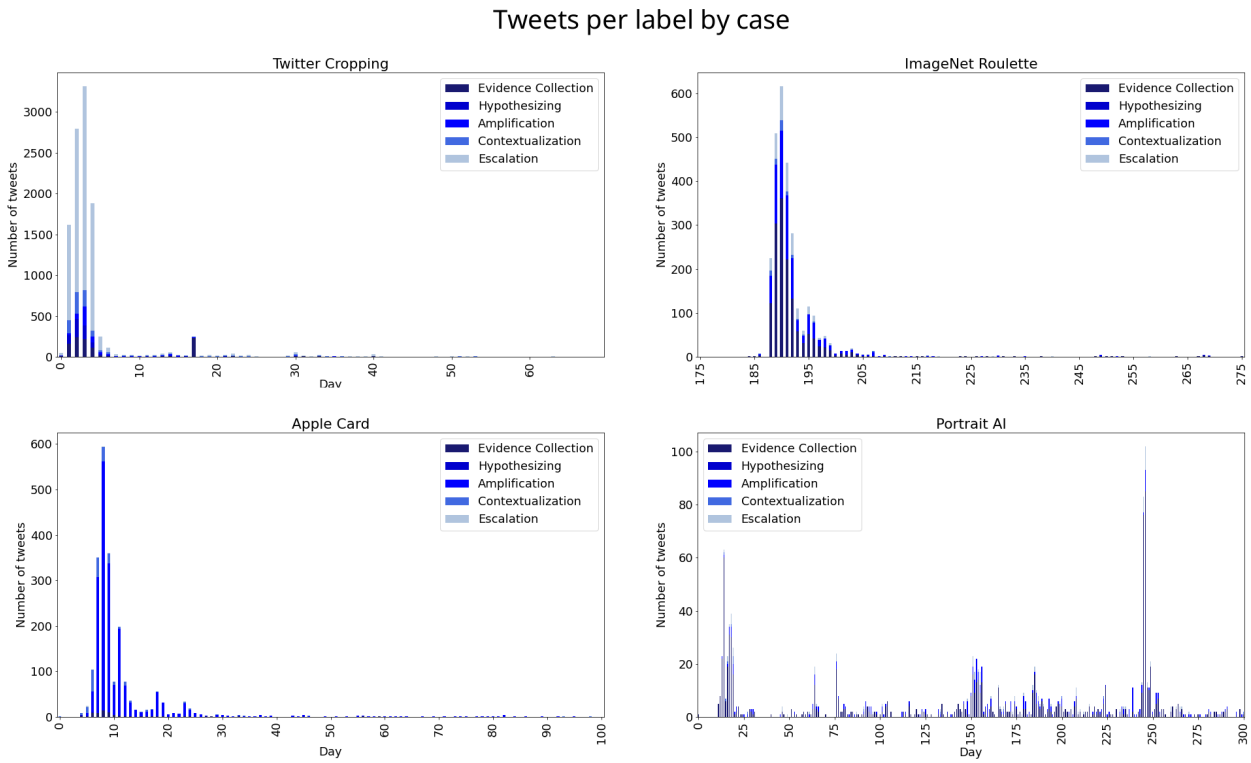


图1：随时间变化的劳动分工／角色推文活动。在所有案例中，横轴均设置为：0 表示发起用户驱动的审计的第 fi一条推文发布之日。柱状图中堆叠的颜色代表扮演5种"劳动分工"角色之一的推文百分比（这些角色列于图例中）。请注意，每个案例的横轴和纵轴各不相同。此外，对于 ImageNet Roulette，关于它的第 fi一条推文发布于其公开发布日前约6个月。因此，出于页面大小和 fi适配性的考虑，该案例的横轴从第175天开始，图 fi展示的是该案例的活动高峰期，该高峰出现在艺术项目公开发布之后。第5.3节将介绍这些图 fi中展示的主要用户审计角色。

表1：四个案例的描述性统计比较。本文将对话定义为以下任一情形：（1）一组共享"conversation_id"的推文；或（2）一条推文及其对应的回复，以及对这些回复的回复。用户贡献定义为参与者在 Twitter 上发布的、与用户驱动的审计相关的推文数量。

	Twitter Cropping	**ImageNet Roulette**	**Portrait AI**	**Apple Card**
日期范围	2020年9月18日 - 2020年11月26日	2019年3月12日 - 2021年5月13日	2020年5月8日 - 2022年1月12日	2019年11月3日 - 2022年1月16日
持续时间	2个月8天	2年2个月1天	1年8个月4天	2年2个月5天
推文总数	11,323	2,813	1,440	2,408
单日推文数最多	3,378	617	102	595
峰值最高日期	2020年9月21日	2019年9月18日	2021年1月9日	2019年11月11日
独立参与者数量	10,160	2,286	1,226	1,687
对话数	883	2,233	1,335	2,071
每位参与者的平均推文数（标准差）	1.11 (2.41)	1.23 (0.78)	1.17（0.73）	1.43（2.49）
中位数	1	1	1	1
每次对话的平均推文数（标准差）	12.82 (129.40)	1.26（1.94）	1.08（0.37）	1.16（4.15）
中位数	1	1	1	1
每次对话的平均参与者人数（标准差）	11.77（124.46）	1.14（1.72）	1.04（0.21）	1.08（2.11）
中位数	1	1	1	1
% 发布 > 1 条推文的参与者	5.87%	13.69%	10.28%	12.66%
点赞总数	695,094	46,698	46,429	200,681
转推总数	10,926	3,807	1,503	8,537

发布了最初的推文，指出该算法存在的问题。在发布这条 fi最初的推文的同一天，Twitter 道歉并概述了将开始采取的措施，以解决该算法的种族偏见问题 [41]。同样，对于 Apple Card，纽约州金融服务局于2019年11月9日报告称，将开始调查高盛的信用评估算法；就在几天前，一名用户发布了原始推文，声称该算法存在偏见 [69]。

相比之下，在所观察的整个时间段内，Portrait AI 案例中的讨论和用户推文活动都呈现出零星的状态。与其他用户驱动的审计相比，自该软件推出以来，只有少数用户对其算法是否存在偏见提出了担忧。此外，平台运营方尚未直接回应这一假设中的偏见问题，也没有任何其他机构介入。我们注意到，与其他三个案例不同，关于 Portrait AI 案例潜在偏见的新闻报道非常少；同时，也没有名人像 Apple Card 案例中那样对此进行讨论。

最后，我们发现，在全部四个案例中，绝大多数参与者都只贡献了一条推文。虽然所有案例都有许多参与者，但按推文数量衡量，真正积极参与的人寥寥无几。此外，尽管这些单条推文中的任何一条可能影响都不大，但我们认为，它们汇总起来提升了这些案例的关注度，并吸引了记者的注意，最终促成了 Twitter 图像裁剪、ImageNet Roulette 和 Apple Card 案例中的变化。

5.2 主要贡献者：他们是谁，又做了什么（RQ2）

本节将介绍我们对各案例主要贡献者的分析结果，重点关注“内容生产者”和“内容传播者”。如第 4 节所述，内容生产者是在某个案例中发布推文最多的人（我们以此作为其对用户驱动审计所作贡献的代理指标）；内容传播者则在点赞和转推等参与度指标上得分最高（我们以此作为传播案例相关信息的代理指标）。我们发现，在积极参与这些审计的推文发布者中，“顶尖内容生产者”（发布推文数量最多的人）与“顶尖传播者”（获得大量点赞和转推的人）之间存在明显分工。换言之，尽管顶尖内容生产者可能贡献了更多推文，但其内容获得的参与度却低于顶尖传播者，这表明用户主导的审计要取得成功，可能需要这两类用户共同参与。

例如，在 ImageNet Roulette、Portrait AI 和 Apple Card 案例中，每个案例都只有一个账户同时跻身“前 10 名制作人”和“前 10 名传播者”。此外，对于 ImageNet Roulette 和 Apple Card 案例，重叠账户恰好也是这两个案例的发起者。在 ImageNet Roulette 案例中，Kate Crawford 被评为第二高产的内容生产者和内容传播者，而 David Heinemeier Hansson 被评为第二高产的内容生产者以及排名最高的内容传播者。

我们还观察到，顶级内容生产者和顶级内容传播者的影响者数量存在差异。我们将“影响者”定义为当前粉丝数超过在我们分析时拥有超过 1,000 名关注者（请注意，这可能与用户参与该案例时的关注者数量不同）。顶级内容传播者拥有的影响者数量多于顶级内容生产者。由于内容传播者是根据其推文收到的点赞和转推数量来界定的，因此 Twitter 上拥有大量关注者的用户会成为顶级内容传播者也就不足为奇了。这表明，获得影响者的关注和参与，有助于用户驱动的审计案例传播。

5.3 用户生成内容中的劳动分工（RQ3）

我们对推文的分析发现，推文在审计案例中发挥了 fi种作用：提出假设、证据收集、扩散、情境化和升级。需要注意的是，我们将与某个案例相关的所有推文都视为“危害识别”的一种形式，因为这些审计正在引起人们对算法偏见的关注（关于我们如何确定推文相关性的详细信息，请参见 4.2）。

表 3 展示了四个审计案例中各类角色的描述及其出现频率。所有角色都出现在所有审计案例中，但各角色的出现频率因案例而异。例如，在 Twitter Cropping 案例中，大多数推文都发挥了升级作用，通过表达对审计过程中所观察到现象的情绪反应来推动升级；而在 Apple Card 案例中，升级推文几乎可以忽略不计。我们将在下文逐一介绍各类角色时，详细讨论这些模式。

为了解各类推文角色在每个案例整个过程中的分布情况，图 1 以堆叠条形图展示了某一天的推文活动中各类角色所占的比例。这表明参与者与角色相互交织，印证了先前研究的发现：用户驱动的算法审计不一定是线性或分阶段的过程 [65]。例如，我们观察到，在 Twitter Cropping、ImageNet Roulette、Portrait AI 和 Apple Card 案例中，用户的推文随着时间推移发挥了每一种角色的作用，即使其中一种角色占据主导地位也是如此。因此，对于这些案例，我们的数据并不表明用户在不同时间发挥不同的角色，也不表明他们在案例早期发挥某些角色、后期则不再发挥这些角色。

5.3.1 提出假设。 第一类推文是提出假设，其中用户提出了他们围绕算法系统形成的不同非正式理论，以解释算法如何运行并产生这些有害结果 [22, 25] 。总体而言，我们观察到两种类型的提出假设类推文：理论假设和实验性假设。在四个案例中，“提出假设”类推文在各自数据集中所占比例均为最小的几类之一（参见表 3）。

理论假设： 在理论假设类推文中，用户陈述了他们认为导致算法偏见的原因。用户对偏见来源的看法，似乎受到其职业或工作经验、知识，以及对算法通常如何进行预测或决策的认识的影响。例如，一位参与者认为，数据集同质性可以解释为什么 ImageNet Roulette 会产生看似带有偏见的图像分fi类：“这值得惊讶吗？不。大多数 #人工智能 都基于

表 2：比较前十名内容生产者和内容传播者。请注意，账户类型的数量可能不等于 10，因为在我们的分析过程中，部分账户已设为私密或被删除。

	Twitter Cropping		ImageNet Roulette		Portrait AI		Apple Card	
	制作人	传播者	制作人	传播者	制作人	传播者	制作人	传播者
# 专家	7	5	6	3	1	2	4	7
# 影响者/名人	4	8	7	9	4	7	8	10
# 个人账号	9	10	8	9	10	7	7	7
# 新闻账号	0	0	0	0	0	0	2	1
# 公司账号	0	0	0	0	0	1	0	1
# 机器人账号	0	0	0	0	0	0	1	0
# 其他	1	0	1	0	0	0	0	0
前十名平均粉丝数	7,068.8	42,460	31,026.9	336,700.2	4,078.7	9,115.9	892,758.8	2,133,274
中位数	839	7,374	11,149	421,517	587.5	4,718	2,720.5	123,054
平均粉丝数（非头部）	95.5	229.5	216.9	236.8	264.1	109.3	211.4	156
中位数	58	127.5	66.5	315.5	179	129.5	87.5	129.5

表 3：劳动分工代码。这些代码描述审计中所做的工作。每条推文最多带有其中一个标签。我们数据集中的具体推文标记为 T#####，例如 T2538。百分比表示应用机器学习模型后，在四个案例中带有此标签的推文所占比例：Twitter 图像裁剪（TC）、ImageNet Roulette（INR）、Portrait AI（PAI）、Apple Card（AC）。

Code	定义	示例	各案例百分比
提出假设	推测所观察到的行为原因的推文 算法正在发生。	"@XXXXX 并不是说这没有问题，但我觉得关键在于……这里发生的情况是，算法更倾向于对比度较高的内容 面部特征与肤色亮度之间。当 转换为二进制代码后，该算法会更加自信fi “浅肤色面孔是一张真实的人脸。”（T2538）	TC：5.69% INR：0.14% PAI：0.49% AC：0.29%
证据收集	包含支持或反对某一观点的证据的推文 其他推文提出的假设通常包括一个 指向一张图片的链接，或者在 Apple Card 案例中，用户……证言/经历。	“我被 ImageNet Roulette 标成了‘小丑、蠢蛋、傻瓜、呆子、快乐的安德鲁’……感觉他们好像认识我一辈子了 [URL Redacted]”（T13072）	TC：9.68% INR：52.44% PAI：72.5% AC：3.65%
放大fi	有助于分享和传播信息的推文，这些信息是其他人参与审计并提升所必需的 对 fi发现及相关讨论的认知。通常 包括标记其他 Twitter 用户和/或分享新闻 有关此事件的文章。	“这是一篇关于围绕 Apple Card 的种种争议的好文章，@XXXXX 该事件涉及 Apple Card 被指明显歧视女性。 原帖读起来像是在抱怨，实在让人恼火—— 但这件事还是有一些值得借鉴之处， 情况如此。 [URL Redacted]”（T16514）	TC：4.66% INR：30.61% PAI：17.57% AC：85.05%
情境化	用于将正在进行的审计置于具体语境中的推文和定位将其置于更广泛的社会、技术或文化语境中。	“@XXX 当科技兄弟试图告诉你，他们为什么不需要上伦理课时”（T10823）	TC：8.09% INR：3.34% PAI：1.11% AC：10.63%
升级	通常是情绪化的推文，主要是对……作出反应的内容。在审计过程中已观察到的现象，也都隐含地增加了—提高案例的可见度并构建一个反公共领域 审计员中的领域。	“Bruuuuuhhhh，PortraitAI 为什么要这样对我？？ [URL 已删节]”（T14284）	TC：69.21% INR：13.30% PAI：8.13% AC：0.33%

关于对许多人的数据进行统计合并。就像一家企业一样。我们需要能够反映人群多样性的人工智能方法，而不是将这种多样性埋没在大数据中。*[URL redacted]*”（*T14034*）。

我们还观察到，用户会提出假设来反驳其他人的观点。例如，另一位用户反驳了上述关于数据集偏差的说法，转而提出，用户缺乏对 ImageNet Roulette 项目初衷的理解：该项目原本是作为一场艺术展览而设计的，旨在展示算法如何存在偏见，而这最能解释为什么用户会观察到偏见：“这篇发人深省的文章探讨了图像识别研究中的政治议题：问题不仅仅在于拥有更好的数据集。（*ImageNet Roulette* 在媒体报道中被过度简化/推销；我认为它是一个更好的项目*fi*”当你明白它是如何夸大一个现实问题的（对于 *effect*。）*[URL 已隐去]*”（*T13876)*。还有人表达了这样的看法：算法偏见源于算法背后存在偏见的人，包括开发人员和研究人员，而不是算法本身：”唯一存在种族和性别偏见的机器学习算法，是那些由“伦理人工智能”研究人员炮制出来的算法。”（*T17807)*。

实验性假设提出：实验性假设推文的特点是，用户提出ff测试策略或行动，供其他人采用，以识别并收集证据，判断算法是否存在偏见。换言之，这些推文建议用户采取进一步行动来支持某个假设。例如，

针对一位用户的测试，另一位参与者建议再进行一些测试，也可以用来评估 Twitter 自动裁剪算法。*“@bascule* 现在试试把 *Obama* 放在上方第一张图片里，再把 *Obama* 放在下方最后一张图片里。也许这只是 *Twitter* 组织图片的方式：*fi*第一张在上，最后一张在下。”（*T2883*）。

我们发现，这些假设的一个特点是，用户提出了不同的理论ff，并且彼此相互影响。提出的理论会纳入和/或重新组合他人关于算法如何运作的看法。例如，在 Twitter 图像裁剪案例中，一些用户假设裁剪掉图像的原因是肤色或种族；另一些用户则提出，可以通过在图像中替换这些特征并观察结果，来检验其他面部特征：*“@XXXXX* 更好的实验方法是人为*fi*加深 *Mitch* 的肤色、提亮 *Obama* 的肤色，再分别测试黑皮肤的 *Mitch* 与白皮肤的 *Mitch*，或黑皮肤的 *Obama* 与白皮肤的 *Obama*。正如其他人指出的，影响算法选择的也可能是其他面部特征，*ff* 例如微笑或眼镜。”（*T2212*）。

5.3.2 证据收集。 第二类内容是证据收集 推文，用户在其中分享自己收集的证据，以记录软件算法的偏见。用户收集的证据通常支持其他用户提出的假设，有时也会推翻其中某项假设。换言之，用户收集的证据使这些审计案例为其他参与者提出的主张提供了数字证据。证据收集类推文也可能为引起权威人士 fi的关注奠定了基础，例如，在 Apple Card 案例中，纽约州金融服务局曾在 Twitter 上作出回应 [14]。

在所有四个案例中，证据收集推文都占各自数据集的相当大比例，例如，ImageNet Roulette 案例中占 52.44%，Portrait AI 案例中占 72.5%。ImageNet Roulette 案例体现了自上而下的用户驱动审计，其中偏见行为已被公开证实，这可能会fl影响参与者，让他们想要亲自测试并观察这种偏见行为，进而可能产生更多“证据收集”推文。Portrait AI 案例在这方面颇为有趣，因为许多人经常分享自己的肖像处理结果，却并未明确尝试审计该系统。

在 Twitter 图像裁剪、ImageNet Roulette 和 Portrait AI 案例中，我们发现，用户收集的数据通常是用户生成、收集并分享到 Twitter 上的修改后的图像：用户将原始图像上传到某个软件应用，该应用的算法随后以某种方式转换图像数据。例如，在推文 15 中，一位用户描述了对图像进行 A/B 测试，以辨别 Twitter 图像裁剪案例中所称的算法偏见是否可被观察到：“好吧，我正在进行一项系统实验，以确认裁剪偏见是否确实存在。我正在通过程序（使用 *tweepy*）发布推文，其中包含一个由以下图像组成的（*3 x 1*）图像网格：一张自我认同为 *fi*黑人男性的图像 + 空白图像 + 一张自我认同为 *fi*白人男性的图像（致谢 *@XXXXX @XXXXX*）”（*T15*）.

不过，在 Apple Card 案例中，用户生成的数据或证据主要由用户生成文本内容构成，例如描述他们申请 Apple Card 经历的推文文本。这些推文记录了用户观察到高盛算法对他们或他人的信贷审批或额度作出不同决策的情况，主要指称该算法存在性别偏见：“真有意思。*@Apple* 仍然存在偏见。我和妻子分别申请了 *Apple Card*。收入和信用评分相同（她的评分更高）。我的额度却高出 *25%*。*@XXXXX” (T15654)*。T17118 中的另一位用户也指控苹果公司和高盛的算法存在种族偏见，称该算法给出的信用度评级低于其他征信机构给出的评级：“这也存在种族偏见。我声明的 *FICO* 分数比其他主要征信机构低了 *100* 分 *#AppleCard* 苹果公司因被指存在性别偏见而面临审查 *[URL Redacted]” (T17118)*。

收集的媒体数据类型及用户对数据的比较：在三个案例中，除 Apple Card 外，我们发现用户通常会测试并分享他人或自己的图像。在 Twitter 图像裁剪案例中，我们发现用户在测试平台的图像裁剪算法时，通常会发布名人、公众人物或政治人物的图像。相比之下，在 ImageNet Roulette 和 Portrait AI 中参与证据收集活动的用户，通常会通过这些软件的算法测试并分享自己的图像。我们观察到，这意味着为 Twitter 图像裁剪案例收集数据的用户所制作和分享的图像可能具有更强的可比性。例如，许多用户测试并分享了巴拉克·奥巴马总统和美国参议员米奇·麦康奈尔的图像裁剪测试结果。相比之下，ImageNet Roulette 和 Portrait AI 的用户通常会测试并分享经过分类的自身图像，通常是自己的面部肖像。

数据收集者将图片发布到 Twitter 后，我们发现用户常常会比较并讨论他们对图片的观察和看法，有时还会提出新的假设，以解释软件的算法为何以给定的方式将图片分类或裁剪。

5.3.3 扩散。 第三类推文是扩*fi*散，用户通过这类推文发布并扩散与审计相关的信息。这些推文既可能直接包含信息与资源，例如其他用户的 fi现，也可能指引新用户获取相关资源。用户发布扩散类推文时，会让他们的关注者了解到正在进行的审计，从而提高案例的可见度，并可能吸引其他人参与审计。扩散类推文在 ImageNet Roulette（30.61%）、Portrait AI（17.57%），尤其是 Apple Card（85.05%）案例中占了很大比例。Apple Card 案例从头到尾都受到大量媒体报道，因此有许多文章可能会被分享。这或许可以解释为何绝大多数 Apple Card 推文都被归类为扩散类推文。ImageNet Roulette 也获得了一些媒体关注，这可能是因为该项目的创作者 Kate Crawford 和 Trevor Paglen 的fl响力。ImageNet Roulette 案例中，证据收集类推文占了很大比例，这或许促使其他 Twitter 用户也亲自尝试。

在每个案例中，我们认为一些用户确有需要向其他人解释审计并提供相关信息。正如各案例中的传播类推文所示，这一角色不同于分享用户所收集数据和证据的推文，因为收集数据的用户并不总会将证据与审计的整体叙事以及相关社会问题联系起来，

[14]https://twitter.com/LindaLacewell/status/1193183785581498369

例如偏见。根据我们的观察，我们认为这些案例中的传播类推文表明，用户也希望获得这类信息。反过来，我们推测，传播类推文可能会促使扮演其他角色的用户创作更多内容，例如收集更多数据。换句话说，发布传播类推文的用户可能会鼓励其他用户在这些审计中承担不同的角色。

最终，我们发现，尽管放大类推文的总量最大，但它们在 Twitter 裁剪案例中所占比例最小。造成这一现象的原因可能有很多，其中一种可能的解释是，Twitter 裁剪案例中的升级类推文（如下所述）占数据集的 69.21%，可能产生了与放大类推文所期望达到的相同效果。下文进一步识别出三类放大类推文：邀请新参与者、链接新闻报道，以及提供信息与资源。

（1）一些用户会标记特定人士，将他们手动加入讨论，并引导潜在的新参与者了解有关审计的信息。例如，在 T3116 中，用户标记了另外两位 Twitter 用户，请他们查看一个汇集 Twitter 自动裁剪算法示例的讨论串：*"@XXXXX @XXXXX @XXXXX 一定要看看这个！"*（*T3116*）。

（2）一些用户会附上报道审计进展的新闻报道，从而扩大相关新闻报道的传播范围和曝光度。例如，针对 Twitter 图像裁剪案例，Twitter 就其算法发布了公开道歉；报纸《卫报》[41] 对此进行了报道，该报道在审计过程中被多次转发：*"Twitter 就其'种族主义'图像裁剪算法致歉 [URL redacted]" (T382)*。

同样，ImageNet Roulette 案例也获得了不少媒体关注，例如《Frieze》杂志 [32]：*"我为 @frieze_杂志撰文谈及 @katecrawford 和 @trevorpaglen 的 #imagenetroulette [URL redacted]" (T13194)*。

2021 年 4 月，Apple Card 针对几个月前在 Twitter 上传播的性别偏见指控推出了一项新功能，用户也将这一更新纳入了审计工作中：*"Apple Card 的新功能旨在应对信贷行业中普遍存在的性别偏见。[URL redacted]" (T15657)*。

（3）有些用户会在回答他人问题时提供信息与资源。通过这种做法，这些用户会向那些对审计相关事件表示兴趣的其他用户提供相关信息与资源，帮助他们在愿意的情况下参与其中。例如，一位 Twitter 用户询问*"@XXXXX 这是什么网站，我想试试" (T12044)*，作为回应，参与者介绍了 ImageNet Roulette：*"@XXXXX Imagenet roulette" (T12048)*。

5.3.4 情境化。 第四类推文是情境化，用户通过这种方式将正在进行的审计置于更广泛的社会、技术或文化语境中，从而拓宽观察范围。他们会提供一般性的社会、技术和文化信息，帮助其他人更好地理解底层算法、有问题的机器行为，以及这些行为为何、如何可能伤害其他社会群体。通过这样做，这些推文有助于为其他人提供信息并指引方向。

可能不具备有效理解审计并参与其中所需的社会背景、技术专长或文化知识的用户。

发布情境化推文的用户可能会批评用户驱动的审计本身，将其他参与者的付出置于更广泛的社会与文化语境中。例如，在下面这条推文中，一位参与者指出，尽管缺乏技术知识，其他参与者仍在提出并测试假设，并将这种行为归因于人性：*"@XXXXX 这个帖子里的所有人都在毫无头绪地探究一个自己并不了解的算法，并根据实验得出不完善的理论。真棒，这就是人类会做的事。有意思，往下翻翻。" (T3212)*。

同样，以下用户对 Apple Card 算法存在偏见的说法表示怀疑，称审计参与者仅凭很少的证据就妄下结论：*"@XXXXX 这是一个很好的例子，说明有人仅凭非常有限的证据就立刻打出'审查！偏见！'这张牌……而（如果你想一想）苹果公司没有理由这样做。任何人都可以用 find DC 以及其他任何方式……这样做只会让人们对苹果公司感到不满" (T15723)*。

情境化推文还可能讨论算法偏见更广泛的背景，从而为审计提供周围的技术语境。如下所示，一位参与者评论了偏见在人工智能中的相关性，以及 ImageNet Roulette 在其中所处的位置：*"@XXXXX 人工智能中的偏见是一个非常有趣且相关的话题。随着我们不断向前发展，它将继续变得更加重要。希望像 #ImageNetRoulette 这样的项目能够提高人们对这一问题的认识，并推动变革。#BC- STT" (T11357)*。

与此类似，以下一位参与者在 Portrait AI 用户驱动的审计 *azine* 中评论了可能影响 Portrait AI 算法开发的技术和文化影响因素：fi响ff：*"@XXXXX @XXXXX @XXXXX 很遗憾，有色人种的古典肖像画并不多。这很糟糕，但我不认为这是开发者的错，他们可能收集了古典艺术家的大量肖像画，并用这些数据训练了人工智能" (T15566)*。

情境化推文在 Apple Card 案例中占比第二高（10.63%），但在其他任何案例中都没有占很大比例。尽管如此，在四个审计案例中，情境化推文都不是占比最低的。这些观察结果可能表明，围绕此类内容创作内容并不流行，或者只有一小部分参与者愿意或能够以这一角色发布推文。由于"情境化"推文需要一定程度的专业知识，因此这类推文可能是由少数参与者创作的。

5.3.5 升级。 fi推文的最后一类是升级。我们发现，很大一部分推文要么是对用户在使用算法时观察到的现象作出的反应，要么是对他人所呈现内容的回应。这些推文带有情绪色彩且具有即时反应性，表达了用户对观察到的算法行为或正在进行的审计的感受。尽管升级推文可能无法为正在进行的审计提供直接证据或帮助形成假设，但它们可能仍提高了审计在 Twitter 等算法中介平台上的可见度，进而增加流量，吸引更多用户参与用户驱动的审计。通过这种方式，这些推文也有助于构建一个

审计员之间的反公共领域 [24]，因为个体参与者现在或许能够围绕共同的话题和／或观点在情感上彼此产生共鸣，从而营造出围绕审计的社群感。这些推文也可能增加与发布审计相关推文的用户互动的用户数量，因为众所周知，具有即时反应性的用户生成内容能够带来这种效果 [14]。

四个案例中的升级推文构成相似。升级推文通常较短，使用首字母缩略词、脏话和俚语，包含表情符号，而且常常带有幽默。这些元素可能有助于增强审计过程中的情绪化和主观影响fl。例如，在下面这条推文中，用户让另一位用户注意到 Twitter 图像裁剪算法对其帖子做了什么，并用大写字母和脏话表达自己的沮丧：*“@XXXXX @XXXXX Twitter 图像裁剪把你坑惨了” (T7450)*。

类似地，ImageNet Roulette 案例中的一位参与者表达了自己对 ImageNet Roulette 如何给其图像分类的看法：*“imagenet roulette 那玩意儿把我整惨了，老兄” (T12894)*。再比如，Apple Card 的一位参与者表达了对其观察到的 Apple Card 算法的沮丧：*“@XXXXX@dhh @AppleCard 这事儿真他妈离谱，这种破事竟然在幕后发生。到底为什么啊。” (T17973)*。相比之下，其他用户则用幽默表达自己的看法：“兄弟，既然能自己给自己画像，为什么还要用 *PortraitAI*？ *” (T14276)*。

升级推文在 Twitter Cropping 案例中占比最高（69.21%），但在另外三个案例中的占比相对较小，尽管目前尚不完全清楚这种差ff异为何存在。此外，在某些情况下，升级推文可以起到放大效fi应，帮助表达对某种算法的负面反应，并提高该案例的可见度。

6 局限性

我们的研究存在若干局限性。本节将介绍这些局限性，以提醒读者在解读研究结果时保持谨慎。我们强烈鼓励在用户驱动的算法审计领域开展更多研究，以弥补并改进下文所述的局限性。

首先，我们仅收集并考察了有关 Twitter 用户驱动审计的数据。其他平台上的用户驱动审计可能具有不同的特征。其次，我们通过在 Twitter 上搜索特定关键词和短语来收集数据，然后使用 Twitter API 和 *snscrape* 自动下载这些 URL 对应的数据。这种以关键词为中心的方法可能遗漏了一些相关推文。不过，使用关键词搜索来收集 Twitter 数据，是研究 Twitter 用户活动的常见方法 [46], , 包括研究用户在 Twitter 上分享有关其他平台用户驱动审计的信息 [44]。

我们的研究还有一项局限：我们只考察了 Twitter 上的 4 个用户驱动审计案例，因此研究结果未必适用于该平台上的其他类型用户驱动审计。例如，偏见的类型和/或算法生成的输出类型、Twitter 平台本身是否适合开展用户驱动审计，以及大众媒体，都可能影响ff谁参与以及如何参与。我们也只考察了推文文本。例如，在 Twitter Cropping 案例中，我们没有分析人们用来对 Twitter 算法进行种族偏见 A/B 测试的图像。此外，我们的一些 Twitter 数据可能由机器人账号生成并放大传播。虽然我们使用了名为 Botometer（https://botometer.osome.iu.edu/）的工具来评估头部内容生产者和广播账户是否为机器人账号，但没有将该工具用于整个数据集。如果某个 Twitter 账户是机器人账号，分享了与用户驱动审计相关的信息，并被纳入我们的数据集，那么将该账户排除在分析之外未必合理，因为这类假设账户也可能 fit我们所描述的角色。最后，我们研究的另一项重要技术局限是：我们呈现了 Twitter 上用户驱动审计参与程度的描述性统计。这些描述性结果无法确fi定地识别造成这些参与程度的原因。

除了这些技术层面的考量，我们研究的另一项局限源于我们作为研究人员的立场。我们对一条推文所起作用的解读，可能与作者和 Twitter 受众的预期信息不同。此外，我们的研究探讨了影响我们研究团队并不属于其中的群体的算法偏见问题，例如 Black Twitter 用户。我们的研究可能以不利于边缘化群体的方式体现或引入自身的偏见。

7 D讨论与设计启示 IONS

本文对 Twitter 上四个备受关注的用户驱动算法审计案例进行了比较分析。基于我们的研究发现，我们现在将讨论用户驱动的算法审计中何为“成功”，如何看待用户在用户驱动审计中的参与，以及未来可用于更好支持此类审计的潜在设计启示。

7.1 用户驱动的算法审计何谓“成功”?

我们的研究引出了一个问题：用户驱动的算法审计中的成功意味着什么？此前关于算法审计的文献提出了一些用于判定审计是否成功的标准。例如，Costanza-Chock 等人指出：“人工智能审计人员依据一套特定标准进行评估，并向受审计方、公众和／或其他参与者（例如监管机构）提供研究发现和建议，或将其作为法律程序中的证据。”然而，尽管我们考察的至少三个案例（Twitter 照片裁剪、ImageNet Roulette 和 Apple Card）都可视为成功，因为它们促成了重大干预，但这些标准并未涵盖我们研究的用户驱动审计的诸多方面。例如，我们考察的用户驱动审计并没有明确的最终目标。各案例的起始推文都提出了一个假设，试图解释相关行为为何发生。然而，没有证据表明，这些起始推文的作者预期其他人会收集数据、检验他们的假设，并提出其他相互竞争的假设。

与传统的 fi一方、二方和第三方审计相比， [17],用户驱动的审计是否成功，或许不再仅仅取决于是否有一套明确的解决方案来fi解决问题。事实上，在本文考察的四项用户驱动的审计中，没有一项为任何实体制定了明确的标准或建议。也没有迹象表明这些审计试图做到全面或完整；其中大部分

数据收集和提出假设都聚焦于相关算法的某个狭窄方面，而非力求全面。有时甚至没有明确的预期结果。下文将介绍一系列不同的视角和标准，它们可以重新定义用户驱动的算法审计中的成功，并为开发实现"成功"所需的平台和流程提供设计启示。

7.1.1 以"关怀伦理"为中心。 思考用户驱动的审计时，一个有用的框架是女性主义伦理学家 Carol Gilligan 对"正义伦理"与"关怀伦理"的区分 [35]。"正义伦理"通常受到占主导地位的社会群体青睐，关注可推广的标准和解决方案。相比之下，"关怀伦理"更关注人际关系以及作为一种美德的关怀。与 Wu 等人 [73] 在一项有关隐私集体行动的参与者研究中发现的情况相似，在我们考察的四个案例中，参与算法集体审计的用户也更倾向于"关怀伦理"，而非"正义伦理"。事实上，我们在此考察的所有案例都由大量聚焦于情感表达（即"升级"和"放大"）的推文构成，而非科学测试、假设和测量。换言之，从"关怀伦理"的视角来看，成功的用户驱动审计或许缺乏明确、普适的规则或解决方案，转而更关注关系、责任和情感表达。因此，未来用户驱动的审计平台或流程在设计时，不应将用户的所有努力都引向科学假设的提出和测试，也应支持情感表达、沟通和社区建设方面的努力。

7.1.2 以偶发、流动的形式促成抵抗。 理解用户驱动审计成功的另一种方式，是将其视为一种"日常抵抗" [65]。这种研究思路并不把抵抗视为高度组织化、旨在发起革命性挑战的行动，而是关注另一种更为偶发、却仍在日常生活中不断挑战既有权力结构的抵抗 [64]。因此，与传统审计相比，用户驱动审计的成功不能仅由其组织结构是否严密来决定。正如我们在四个案例中所见，参与用户驱动审计的人可以随时加入或离开，并按照自己的意愿投入多少精力。审计活动也缺乏严密组织，许多人都主动自行测试。三个审计案例（Twitter 图像裁剪、Apple Card 和 ImageNet Roulette）的持续时间都较短，各案例的大部分推文都集中在短短几天内发布。然而，这种高度自发且流动的组织形式，也让用户在日常生活中能够根据自身的舒适程度、时间和实际条件（例如交通、互联网接入、照护子女等）自由参与 [59]。对于设计者而言，这表明需要审慎设计引导用户的支架式方法，同时保留一定程度的自发性和灵活性，为用户提供多种参与选择。

7.1.3 形成微型反公众。 用户驱动的审计也可以从"反公众" [31], 的视角来理解，也就是说，将其视为一种社区建设的努ff力。正如我们在全部四个案例中所看到的，尽管大多数用户只贡献了少量推文，但他们仍然围绕新出现的、有问题的机器行为形成了临时的小型在线社群。我们可以将这类社群视为南希·弗雷泽所称的"反公众" [31] ——通常来自边缘化社会群体的成员共同参与属于他们自己的意义建构、意见形成和共识构建。因此，用户驱动的审计取得成功，也意味着用户聚集起来构建微型反公众，作为社区建设努ff力的一部分。这要求在用户驱动的算法审计平台和流程中设计并构建相应的功ff能，使用户能够在这些平台和流程的支持下聚集、讨论并建设社群。

7.1.4 提高公众意识。 最后，与其他传统审计不同，后者的最终目标是实现明确的技术补救；我们发现，"提高公众意识"是用户驱动审计取得成功的关键维度 [65]。近期文献指出，如果系统开发者不愿在平台上托管用户审计工具，"终端用户审计的主要目的将是通过指出并引起人们对有问题的系统行为的关注来促成改变" [49]。事实上，尽管我们在此考察的四个案例都没有明确的解决方案或结构完善的行动计划，但通过"提高公众意识"的过程，它们仍然带来了不同程度的改变：公开分享假设和证据，并通过扩大传播和升级，积极提升审计的可见度。在 Twitter 图像裁剪案例中，Twitter 迅速承认了失误，解释了他们如何测试算法及其方法的不足，并说明了应对该问题的一些后续步骤 [3]。在 Apple Card 案例中，纽约州金融服务局很快表示将展开调查 [69]。对于 ImageNet Roulette，虽然没有立即采取行动，但最终还是从数据集中移除了恶劣的标签 [62]。

7.2 用户驱动的算法审计中的参与：行动主义还是懒人行动主义?

正如先前研究所述，在许多情况下，算法审计可被视为一种行动主义，"一种直接行动实践，其效果往往是引起人们对某一议题的关注，并推动政治与社会变革" [52]。然而，在本研究中，绝大多数参与者只贡献了一条推文，而且大多是以"升级"推文的形式。有些人可能会将这种行为描述为"懒人行动主义"，也就是一种"低风险、低成本的活动 [...]，其目的是提高认知、促成改变，或让参与活动的人获得满足感" [61]。这种行为通常与"务实行动主义"形成对比，后者是"为实现社会变革而采取直接、积极且往往具有对抗性的行动" [61]。因此，这里引出的问题是：我们应如何理解这些"懒人行动主义者"在用户驱动的算法审计中的作用?

我们认为，尽管在更传统的算法审计中，目标是以系统化的方式揭示算法系统中的偏见，因此这些"懒人行动主义者"及其单条推文看起来可能无济于事，但在用户驱动的审计中，大量懒人行动主义者的推文实际上可能产生显著影响。一方面，他们的推文表明，他们对所感知到的有害行为感到强烈的惊讶或愤怒。另一方面，他们的推文还

有助于提高人们对该案例的认识（此前的研究也讨论过这一点[51, 61]），他们通过与关注者分享该案例，并借助算法将其推上 Twitter 热门话题。我们还认为，这两个因素有助于引发足够的担忧，从而引起其他实体的关注，使其能够进一步提高人们对该案例的认识（例如，大众媒体），或直接干预有害的机器行为（例如，监管机构或使用该算法的公司）。这与Cabrera等人的观点一致：“有时，在数字时代，行动主义依赖于大众的懒人行动主义”[12]。但是，懒人行动主义者如何参与审计过程，推动改变呢?

我们观察到，在这些审计案例中，“影响者”和“专家”是主要贡献者，他们在引起人们对问题的关注，以及可能吸引更多人参与方面发挥了关键作用。苹果公司 Apple Card 案例中的影响者尤其引人注目，一些名人通过自己的关注者扩大了该案例的影响力。影响者很可能是用户驱动的审计取得成功的关键，他们能够大幅提高公众意识，并帮助招募潜在参与者。此外，在许多案例中，“专家”——其中许多人担任内容生产者和内容传播者——也通过提出假设、反假设、情境化内容和数据，积极引导审计的方向。这可能部分源于他们对该问题的共同兴趣，以及其技术能力或数据分析技能。因此，尽管大多数用户驱动的审计并非由集中式组织者领导，这些“影响者”和“专家”仍发挥着“领导者”的作用，帮助传播消息并引导其他用户的努力。

懒人行动主义者、影响fl者和专家的参与及其复杂的互动关系，要求我们谨慎考虑设计方案，以便让不同类型的用户有效ff参与用户驱动的算法审计。这包括：a）为懒人行动主义者提供便捷、快速且可随时参与的方式；b）通过提供赏金等激励机制，以及反馈其参与所产生的影响，鼓励影响fl者和专家参与审计过程；c）设计“领导者”（例如影响fl者和专家）与懒人行动主义者之间的沟通渠道，以更好地引导审计工作ff。

7.3 面向用户参与算法审计的设计

我们的长期目标之一是设计专门帮助用户审计其他平台的工具。基于这一背景，我们将探讨支持用户驱动审计的工具所面临的一些设计问题。

7.3.1 面向讨论与审议的设计。 我们的研究表明，用户通过参与对话串来开展算法审计，并借此在其他人的假设、想法和讨论的基础上进一步展开探讨。然而，表1显示，每个案例都有数百条独立的对话，这强烈表明，这些对话串大多是碎片化且分散的。也就是说，一位用户可能会发布一条推文，评论某个案例中可能存在的偏见，并附上相关证据。第二天，另一位用户可能会独立做出同样的事情，却没有意识到其他人已经发现了同一个问题。Twitter 的话题标签在这里并不足以解决问题，因为这要求人们使用相同的话题标签。这种碎片化使用户难以查看所有提出的假设，以及迄今收集的所有证据。

对设计的一项启示是，任何用于支持用户驱动审计的工具，都应考虑如何让人们能够轻松地在同一处收集数据、查看不同的假设、支持或反驳这些假设的证据，以及相关讨论。

然而，另一个相关挑战是数据量庞大。正如我们所见，有些案例中有数千条推文。将这些内容汇总到一处，有助于人们了解案例的来龙去脉，但也会让人难以单独查看证据等内容。此外，某些类型的算法偏见恶劣至极，一个例子就足以说明问题（例如，计算机视觉系统将黑人标记为“猩猩”）；而许多其他类型的算法偏见，则可能需要进行多次比较才能得出结论。例如，仅凭一两个 Twitter 照片裁剪或 Apple Card 案例，很难得出结论；但有了多个例子，就能显现出清晰的规律。因此，另一项设计启示是，工具应便于同时查看多个证据实例，从而更容易进行比较并发现潜在规律。

7.3.2 应对访问／可见性障碍。 用户驱动的审计面临的一项挑战是，用户对算法系统缺乏足够的访问权限／可见性，且难以进行测试，这可能会拖慢或阻碍审计。无法访问服务是审计普遍面临的已知问题[17],，但对用户驱动的审计而言，这个问题尤为突出，因为“并非所有算法行为都同样可见”[23],，因此用户也无法同样地对其进行测试。例如，参与者很容易测试 Twitter 图像裁剪、ImageNet Roulette 和 Portrait AI 案例。这些软件广泛可用，而且人们很容易评估其偏见。然而，Apple Card 案例受到限制，因为申请人必须已婚，而且还必须申请这张信用卡，这可能会影响个人的信用评级。对于其他系统，受影响用户甚至可能无法使用相关服务；例如，在许多高风险算法决策系统中（如住房评估工具、儿童福利预测系统和预测性警务系统），受这些系统影响最大的人群（如边缘化群体）无法接触（有时甚至不知道其存在）相关算法流程，更遑论测试和审计该系统[23]。

这一系列访问障碍呼唤相应的设计考量与流程，让用户更清楚地了解不透明算法系统的存在及其运作方式。例如，Eslami 等人开发了一款工具，让 Facebook 用户能够看到哪些好友的帖子被 Facebook 的信息流策展算法过滤[26],，从而更容易评估和测试该算法。另一个例子是，为了帮助发现范围更广的问题，Twitter 通过偏见悬赏挑战[15]将其图像裁剪代码开源。这不仅发现了此前已检测出的种族偏见，还发现了其他一些问题。

这种额外的可见性，无论来自平台本身还是第三方实体，都有助于用户更好地了解算法系统的运行机制，从而更有效地对其进行测试。此外，这种可见性和访问权限能够吸引人们的

注意力，促使他们参与其中并与他人分享。具体而言，既有研究 [66] 表明，情绪激烈的 Twitter 消息更有可能被转发，而导致愤怒的有害偏见也更容易传播，并吸引更多人加入审计。

测试的一种潜在互补方法是获取底层数据，以便参与者检查大量结果。然而，如果未能建立适当的政策结构，这种方法可能会因ff而失效。例如，在 Apple Card 案件中，纽约州金融服务部最终认定，高盛算法不存在观察到的偏见。作为回应，我们的团队 fi信息自由申请，从中得知该州遗失了调查 fi案，且其保留的记录不包含任何表明申请人性别的数据（回复信见附录）。纽约州还将我们的公开记录申请通知了高盛。作为回应，这家投资银行要求继续拒绝公众查阅调查记录。[15] 这次未成功的申请表明，有必要建立适当的政策和监管结构，以开展算法审计并落实问责；我们将在下一节讨论这些内容。

7.3.3 制定算法审计与问责的政策和法规。 如上一节所示，提高算法系统的可见性／访问权限并非总是易事。组织通常会反对提高算法系统的可见性／透明度，此外，向用户提供可见性／透明度本身也面临固有挑战 [4]。

监管机构是一个可能的着力点。也就是说，要求提供这些算法服务的组织承担更多问责，并更加开放，有助于促进用户参与用户驱动的算法审计，并为用户创造更多机会，以大规模发现和报告偏见。监管机构制定政策时，也应涵盖发现偏见之后的各个阶段。如果组织旨在建立合法性并赋能用户，就需要公开其改进已报告问题所遵循的流程。Twitter 向用户说明其图像裁剪算法存在种族偏见，是一个很好的例子 [15]。

用户驱动审计面临的一个最终问题是组织的回应。具体来说，Twitter 对图像裁剪问题的回应是移除这项功能。其他公司在人们发现有害偏见后也采取了类似的应对措施。例如，在 Buolamwini 和 Gebru 的研究显示这些服务对黑人，尤其是黑人女性，出现更多错误后，微软、亚马逊和 IBM 都暂停了人脸识别服务的访问权限 [10]。在辅助技术领域，Bennett 等人描述了视力障碍者如何从中获益良多 fi，但也对计算机视觉系统可能带来的偏见、错误和歧视表示担忧，因为这类系统可能会描述图片中人物的种族、性别和残障状况 [5]。然而，尽管可能带来诸多益处 fi，开发这套计算机视觉系统的公司还是决定不部署它，很可能是出于对负面舆论的担忧。一方面，在上述每个例子中，有害偏见都被消除。另一方面，潜在有用服务的访问也被阻止了；如果某些情况下特定类型的错误可以容忍或缓解，这可能会带来问题。总之，公司应如何应对有害偏见并不简单，研究人员、伦理学家、政策制定者和从业者值得对此展开更多讨论。

[15]这家投资银行声称，作者——大学研究人员——对其利润构成了严重的竞争威胁，因为他们能够获取该算法用于信贷判定的部分数据。

8 结论

本文对 Twitter 上四个备受关注的用户驱动审计案例进行了比较研究。我们的分析表明，这些案例呈现出相似的用户参与模式：活动量先大幅激增，随后进入低活跃度的长尾阶段；而且大多数用户只贡献了一条推文。我们还发现，“内容贡献者”与“内容传播者”之间存在明确分工，这表明这两种角色对于开展用户驱动的审计都至关重要。最后，我们观察到，这四个案例都包含相似类型的用户生成内容，包括提出假设、收集证据、扩大传播、情境化和升级。我们的发现揭示了 Twitter 上用户驱动审计的若干独特参与模式，并为未来如何谨慎地支持这些自发且灵活的集体行动提供了启示。关于未来研究，一个方向是开发工具，促进开展此类用户驱动审计。我们观察到，劳动分工中存在若干具体角色。如何更有效地实现这种劳动分工？如何激励足够多的人以及足够多元的群体参与？我们如何支持这些参与者收集数据、提出假设并分析结果？其中一些问题可能需要采用其他研究方法来回答，例如深度访谈。我们将这些问题留待未来研究。

参考文献

[1] [无日期]。“苹果公司 *Apple Card* 今日面向所有美国客户推出”。 https://www.apple.com/newsroom/2019/08/apple-card-launches-today-for-all-us-customers/ [2] 2020。Twitter 调查图像预览中的种族偏见。*BBC*新闻 （2020年9月）。 https://www.bbc.com/news/technology-54234822 [3] 2020 Parag Agrawal 和 Dantley Davis。《图像裁剪透明度及即将进行的变更》。*Twitter* （2020年）。 https://blog.twitter.com/en_us/topics/product/2020/transparency-image-cropping.html [4] Mike Ananny 和 Kate Crawford。2018年。《看见却不知晓：透明度理想的局限及其在算法问责中的应用》。《新媒体与社会》20卷3期（2018年），973–989页。[5] Cynthia L Bennett、Cole Gleason、Morgan Klaus Scheuerman、Jeffrey P Bigham、Anhong Guo 和 Alexandra To。2021年。“这很复杂”：协商图像描述中的可及性与种族、性别及残障的（错误）表征。载于*2021*年*CHI*计算机系统人因会议论文集。1–19页。[6] Reuben Binns、Max Van Kleek、Michael Veale、Ulrik Lyngs、Jun Zhao 和 Nigel Shadbolt。2018年。“这把一个人简化成一个百分比”；对算法决策中正义的看法。*CoRR* abs/1801.10408（2018年）。arXiv:1801.10408 http://arxiv.org/abs/1801.10408。Tolga Bolukbasi、Kai-Wei Chang、James Y Zou、[7] Venkatesh Saligrama 和 Adam T Kalai。2016年。男人之于计算机程序员，犹如女人之于家庭主妇？词嵌入去偏。神经信息处理系统进展29（2016年），4349–4357页。[8] Virginia Braun 和 Victoria Clarke。2006年。在心理学中运用主题分析。《心理学定性研究》3卷2期（2006年），77–101页。[9] Anna Brown、Alexandra Chouldechova、Emily Putnam-Hornstein、Andrew Tobin 和 Rhema Vaithianathan。2019年。迈向公共服务中的算法问责：受影响社群对儿童福利服务中算法决策的看法之定性研究。载于*2019*年*CHI*计算机系统人因会议论文集。1–12页。[10] Joy Buolamwini 和 Timnit Gebru。2018年。Gender Shades：商业性别分fi类中的交叉性准确率差异。载于*FAT*。[11] Ángel Alexander Cabrera、Abraham J Druck、Jason I Hong 和 Adam Perer。2021年。通过众包故障报告发现并验证人工智能错误。*ACM*人机交互论文集 5卷CSCW2（2021年），1–22页。

[12] Nolan L Cabrera、Cheryl E Matias 和 Roberto Montoya。2017年。行动主义还是懒人行动主义？社交媒体在当代学生行动主义中的潜力与陷阱。《高等教育多样性杂志》10卷4期（2017年），400页。[13] Zhilong Chen、Hancheng Cao、Fengli Xu、Mengjie Cheng、Tao Wang 和 Yong Li。2020年。理解中介者在在线社交电子商务中的作用：对贝店的探索性研究。《ACM人机交互论文集》4卷CSCW2（2020年），1–24页。[14] Myoung-Gi Chon 和 Hyojung Park。2020年。数字时代的社交媒体行动主义：对争议议题行动主义整合模型的检验。《新闻与大众传播季刊》97卷1期（2020年），72–97页。[15] Rumman Chowdhury 和 Jutta Williams。2021年。推出 Twitter 首个算法偏见悬赏挑战。Twitter（2021年）。https://blog.twitter.com/ engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge[16] Seth Cooper、Firas Khatib、Adrien Treuille、Janos Barbero、Jeehyung Lee、Michael Beenen、Andrew Leaver-Fay、David Baker、Zoran Popović 等。2010年。通过多人在线游戏预测蛋白质结构。《Nature》466卷7307期（2010年），756–760页。[17] Sasha Costanza-Chock、Inioluwa Deborah Raji 和 Joy Buolamwini。2022年。谁来审计审计员？算法审计生态系统实地扫描的建议。载于《2022年ACM公平性、问责与透明度会议论文集》（韩国首尔）（FAccT ’22）。美国计算机协会，纽约，美国，1571–1583页。https://doi.org/10.1145/3531146.3533213[18] Justin Cranshaw 和 Aniket Kittur。2011年。博学者项目：数学领域成功在线协作的经验教训。载于SIGCHI计算机系统人因会议论文集。1865–1874页。[19] Kate Crawford 和 Trevor Paglen。2021年。挖掘人工智能：机器学习训练集中的图像政治。《人工智能与社会》（2021年），1–12页。[20] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。2009年。ImageNet：大型层级图像数据库。载于2009年IEEE计算机视觉与模式识别会议论文集。IEEE，248–255页。[21] Michael A. DeVito、Jeremy Birnholtz、Jeffery T. Hancock、Megan French 和 Sunny Liu。2018年。人们如何形成关于社交媒体信息流的民间理论，以及这对我们研究自我呈现意味着什么。载于2018年CHI计算机系统人因会议论文集（加拿大魁北克省蒙特利尔）（CHI ’18）。美国计算机协会，纽约，美国，1–12页。https://doi.org/10.1145/3173574.3173694。/。[22] Michael A DeVito、Darren Gergle 和 Jeremy Birnholtz。2017年。“算法毁了一切”# RIPTwitter、民间理论，以及对社交媒体算法变更的抵制。载于2017年CHI计算机系统人因会议论文集。3163–3174页。[23] Alicia DeVos、Aditi Dhabalia、Hong Shen、Kenneth Holstein 和 Motahhare Eslami。2022年。迈向用户驱动的算法审计：探究用户揭示有害算法行为的策略。载于CHI计算机系统人因会议论文集。1–19页。[24] John Downey 和 Natalie Fenton。2003年。新媒体、反公共宣传与公共领域。《新媒体与社会》5卷2期（2003年），185–202页。[25] Motahhare Eslami、Karrie Karahalios、Christian Sandvig、Kristen Vaccaro、Aimee Rickman、Kevin Hamilton 和 Alex Kirlik。2016年。一开始我“喜欢”它，后来我把它隐藏起来：社交信息流的民间理论。载于2016年CHI计算机系统人因会议论文集。2371–2382页。[26] Motahhare Eslami、Aimee Rickman、Kristen Vaccaro、Amirhossein Aleyasen、Andy Vuong、Karrie Karahalios、Kevin Hamilton 和 Christian Sandvig。2015年。“我一直以为自己和 [她]并没有那么亲近”：探讨新闻信息流中的不可见算法。载于第33届ACM计算机系统人因年会论文集。153–162页。[27] Motahhare Eslami、Kristen Vaccaro、Karrie Karahalios 和 Kevin Hamilton。2017年。“当心，情况可能比表面看起来更糟”：理解评分平台中的有偏算法及用户对此的行为。载于国际AAAI网络与社交媒体会议论文集，第11卷。[28] Motahhare Eslami、Kristen Vaccaro、Min Kyung Lee、Amit Elazari Bar On、Eric Gilbert 和 Karrie Karahalios。2019年。用户对在线评论平台中算法不透明性与透明度的态度。载于2019年CHI计算机系统人因会议论文集。1–14页。[29] Casey Fiesler 和 Nicholas Proferes。2018年。“参与者”对 Twitter 研究伦理的看法。《社交媒体与社会》4卷1期（2018年），205630511876336。https: //doi.org/10.1177/2056305118763366。/ Kristie Fisher、Scott Counts 和 Aniket Kittur。2012年。分布式意义建构：利用先前用户的努力改进意义建构。载于2012年CHI计算机系统人因会议论文集。247–256页。[31] Nancy Fraser。2021年。重新思考公共领域：对现实存在的民主之批判的贡献。载于《公共空间读本》。Routledge，34–41页。[32] Orit Gat。2019年。“人工智能自拍应用 ImageNet Roulette 如何席卷互联网”。（2019年）。https://www.frieze.com/article/how-ai-selfie-app-imagenet-roulette-took-internet-storm[33] R Stuart Geiger 和 Aaron Halfaker。2013年。利用编辑会话衡量 Wikipedia 中的参与度。载于2013年计算机支持的协同工作会议论文集。861–870页。[34] Susan A Gelman 和 Cristine H Legare。2011。《概念与民间理论》。《人类学年度评论》40（2011），379–398。[35] 卡罗尔·吉利根。1993。《另一种声音：心理学理论与女性发展》。哈佛大学出版社。[36] Nina Grgic-Hlaca、Elissa M. Redmiles、Krishna P. Gummadi 和 Adrian Weller。2018。《人们对算法决策公平性的认知：刑事风险预测案例研究》。（2018），903–912。https://doi.org/10.1145/3178876.3186138[37] Gabriel Grill 和 Nazanin Andalibi。2022。《数据主体对情绪识别透明度和准确率的态度与民间理论》。《ACM人机交互汇刊》6，计算机支持的协同工作1，第78篇文章（2022年4月），35页。https://doi.org/10.1145/3512925。/ Aaron Halfaker 和 R Stuart Geiger。2020。《Ores：通过维基百科中的参与式机器学习降低门槛》。《ACM人机交互汇刊》4，CSCW2（2020），1–37。[39] Aaron Halfaker、Aniket Kittur 和 John Riedl。2011。《别欺负新手：撤销操作如何影响维基百科工作的数量和质量》。载于《第七届维基与开放协作国际研讨会论文集》。163–172。[40] Aniko Hannak、Gary Soeller、David Lazer、Alan Mislove 和 Christo Wilson。2014。《衡量电子商务网站上的价格歧视与导向操控》。载于《2014年互联网测量会议论文集》。305–318。[41] Alex Hern。2020。《Twitter就“种族主义”图像裁剪算法道歉》。《卫报》（2020年9月）。https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm[42] Khari Johnson。2020。《Twitter照片算法被发现存在明显的种族偏见》。（2020年9月）。https://venturebeat.com/2020/09/20/apparent-racial-bias-found-in-twitter-photo-algorithm/[43] Nadia Karizat、Dan Delmonaco、Motahhare Eslami 和 Nazanin Andalibi。2021。《算法民间理论与身份：TikTok用户如何共同生产身份知识并参与算法抵抗》。《ACM人机交互汇刊》5，计算机支持的协同工作2，第305篇文章（2021年10月），44页。https://doi.org/10.1145/3476046[44] Sara Kingsley、Proteeti Sinha、Clara Wang、Motahhare Eslami 和 Jason I. Hong。2022。《“给每个人 [……] 多一点公平”：内容创作者对与弱势群体相关内容遭算法取消变现的看法与回应》。《ACM人机交互汇刊》6，计算机支持的协同工作2，第424篇文章（2022年11月），37页。https://doi.org/10.1145/3555149。/ Sara Kingsley、Clara Wang、Alex Mikhalenko、Proteeti Sinha 和 Chinmay Kulkarni。2020。《审计数字平台中的经济机会广告歧视》。https://doi.org/10.48550/ARXIV.2008.09656。/ Shamika Klassen。2022。《黑人Twitter就是宝藏：为何这个线上社群值得研究，以及如何以尊重的方式开展研究》。《Interactions》29，1（2022年1月），96–98。https://doi.org/10.1145/3505681。/ Will Knight。2019。《“苹果公司的 Apple Card 没有‘看到’性别——这正是问题所在”》。（2019年）。https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/[48] Irene Font Peradejordi Kyra Yee。2021。《“分享首届算法偏见悬赏挑战的经验”》。Twitter（2021）。https: //blog.twitter.com/engineering/en_us/topics/insights/2021/learnings-from-the-first-algorithmic-bias-bounty-challenge[49] Michelle S. Lam、Mitchell L. Gordon、Danaë Metaxa、Jeffrey T. Hancock、James A. Landay 和 Michael S. Bernstein。2022。《终端用户审计：赋能社群主导大规模有害算法行为调查的系统》。《ACM人机交互汇刊》6，计算机支持的协同工作2，第512篇文章（2022年11月），34页。https://doi.org/10.1145/3555625。/ J Richard Landis 和 Gary G Koch。1977。《分类数据观察者一致性的测量》。《生物统计学》（1977），159–174。[51] Yu-Hao Lee 和 Gary Hsieh。2013。《懒人行动主义会损害行动主义吗?道德平衡与一致性对线上行动主义的影响》。载于《SIGCHI计算机系统人因会议论文集》。811–820。[52] Danaë Metaxa、Joon Sung Park、Ronald E Robertson、Karrie Karahalios、Christo Wilson、Jeff Hancock、Christian Sandvig 等。2021。《算法审计：从外部向内理解算法系统》。《人机交互基础与趋势》14，4（2021），272–344。[53] Carolina Milanesi。2021。《“Apple Card的新功能应对信贷领域最大的难题之一”》。（2021年）。https://www.fastcompany.com/90627796/apple-card-apple-card-family-share-card-children[54] Safiya Umoja Noble。2018。《压迫算法：搜索引擎如何强化种族主义》。（2018年1月9日）。https://www.jstor.org/stable/j.ctt pwt w[55] 纽约州金融服务部。2021。《“纽约州金融服务部关于 Apple Card 调查的报告”》。（2021年）。https://www.dfs.ny.gov/system/files/documents/ 2021/03/rpt_202103_apple_card_investigation.pdf[56] Parmy Olson。2018。《助长谷歌翻译性别歧视的算法》。《福布斯》（2018年2月）。https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/[57] Charles C. Ragin 和 Howard S. Becker。1992。《什么是案例？：探索社会调查的基础》。剑桥大学出版社。

[58] Inioluwa Deborah Raji 和 Joy Buolamwini。2019。可执行审计：调查公开点名商业人工智能产品有偏差的性能结果所产生的影响。载于*2019*年*AAAI/ACM*人工智能、伦理与社会大会论文集（美国夏威夷州檀香山）（*AIES '19*）。美国计算机协会，美国纽约州纽约，429 435 10 1145 3306618 3314244 – 。https://doi.org/ 。 / 。 [59] Brandon Reynante、Steven P Dow 和 Narges Mahyar。2021。。开放式公民设计框架：整合公众参与、众包和设计思维。数字政府：研究与实践2，4（2021），1–22。[60] Ronald E. Robertson、Shan Jiang、Kenneth Joseph、Lisa Friedland、David Lazer 和 Christo Wilson。2018。在 Google 搜索中审计算法的党派受众偏差。*ACM*人机交互会议论文集2，CSCW，文章148（2018年11月），22页。https://doi.org/10.1145/3274417[61] Dana Rotman、Sarah Vieweg、Sarita Yardi、Ed Chi、Jenny Preece、Ben Shneiderman、Peter Pirolli 和 Tom Glaisyer。2011。从懒人行动主义到行动主义：社交媒体时代的参与式文化。载于*CHI'11*计算系统中的人为因素扩展摘要。819–822。[62] Christina Ruiz。2019。艺术项目揭露其种族主义偏见后，领先的在线数据库将删除60万张图片。《艺术新闻》（2019），23。[63] Christian Sandvig、Kevin Hamilton、Karrie Karahalios 和 Cedric Langbort。2014。算法审计：检测互联网平台歧视的研究方法。数据与歧视：将关键担忧转化为富有成效的探究22（2014），4349–4357。[64] James C Scott。1985。弱者的武器：农民抵抗的日常形式。耶鲁大学出版社。[65] Hong Shen、Alicia DeVos、Motahhare Eslami 和 Kenneth Holstein。2021。普通用户的算法审计：理解普通用户揭示有害算法行为的力量。*ACM*人机交互会议论文集5，CSCW 2，433 2021 29 10 1145 3479577 文章（10月），页码。https://doi.org/ 。 / Stefan Stieglitz 和 Linh Dang-Xuan。2013。。社交媒体中的情绪与信息扩散——微博的情感倾向和分享行为。《管理信息系统杂志》29，4（2013），217–248。[67] Latanya Sweeney。2013 56，5 2013 44 54 。在线广告投放中的歧视。《*ACM*通讯》（），– 。[68] Niels van Berkel、Jorge Goncalves、Danula Hettiachchi、Senuri Wijenayake、Ryan M. Kelly 和 Vassilis Kostakos。2019。众包对机器学习公平预测器的认知：再犯案例研究。*ACM*人机交互会议论文集3，CSCW，文章28（2019年11月），21页。https: //doi.org/ 10 1145 3359130 . / Neil Vigdor。。苹果公司在遭遇性别歧视投诉后调查 Apple Card。（2019）。[69] 2019。https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html [70] Luis Von Ahn 和 Laura Dabbish。2004。用电脑游戏为图像添加标签。载于*SIGCHI*计算机系统人因大会论文集。319–326。[71] Karl E Weick。1995。组织中的意义建构。第3卷。Sage。[72] Allison Woodruff、Sarah E. Fox、Steven Rousso-Schindler 和 Jeffrey Warshaw。2018。对算法公平性认知的定性探索。（2018），1 14 10 1145 3173574 3174230 – 。https://doi.org/ 。 / 。Yuxi Wu、W Keith Edwards 和 Sauvik Das。2022。“提出这样的要求合情合理”：迈向隐私集体行动中的统一声音。载于*CHI*计算机系统人因会议。1–17。[73] [74] Kyra Yee、Uthaipon Tantipongpipat 和 Shubhanshu Mishra。2021。Twitter上的图像裁剪：公平性指标及其局限，以及表征、设计与能动性的重要性。*ACM*人机交互会议论文集5，CSCW2，文章450 2021 24 10 1145 3479594 （10月），页码。https://doi.org/ 。 / Cat Zakrzewski。。科技：民主党人提出禁止算法歧视的法案。《华盛顿邮报》（2021）。[75] 202

一封信息自由法（FOIL）申请信

本函来自纽约州金融服务部，旨在回应我们关于 Apple Card 案件的数据请求。尽管该函是由本次投稿的一位共同作者提出请求的，但回函的收件人却是高盛的律师。由于该函属于纽约州政府制作的公共记录，因此我们保留了原函中的所有姓名信息，而未对其进行删减（在同行评审期间，提出该请求的共同作者姓名曾被隐去）。

KATHY HOCHUL
Governor

ADRIENNE A. HARRIS
Superintendent

VIA EMAIL
(MDenerstein@gibsondunn.com)

June 2, 2022

Mylan L. Denerstein
Gibson Dunn & Crutcher LLP
200 Park Avenue
New York, NY 10166-0193

Re: Freedom of Information Law (“FOIL”) Request - Tracking No. 2022-089785

Dear Ms. Denerstein:

Pursuant to N.Y. Public Officers Law § 89(5)(b)(3), this correspondence represents the determination of the New York State Department of Financial Services (“Department”) of the request by your client, Goldman Sachs Bank USA (“Goldman”), that records it submitted to the Department continue to be excepted from FOIL disclosure under N.Y. Public Officers Law § 87(2)(d). For the reasons set forth below, your client’s request is granted.

I. Background

The Department received the following FOIL request on January 25, 2022:

> Dear NYSDFS, We request a copy of the data in this report: https://www.dfs.ny.gov/system/files/documents/2021/03/rpt_202103_apple_card_investigation.pdf Specifically, we request the records for the 400,000 applicants, including these variables: gender, race, age, salary or income, credit score, credit limit and if possible zip code.

The Department construes this request to seek records containing data of applicants submitted to and reviewed by the Department during its investigation of Goldman Sachs’ underwriting of Apple-branded credit cards, which includes the gender, race, age, salary, income, credit score, credit limit and zip codes of applicants for such card. After a diligent search of the Department's records, a record containing all such data could not be located.

II. Trade Secret Determination

Notwithstanding that determination, the Department has located a record that contains some of the demographics mentioned in the request, which is a spreadsheet of applicant data that includes a) application identification numbers, b) applicant names, c) applicant dates of birth, d) FICO scores, e) zip codes, and f) application decisions (“Applicant Data”). When Goldman submitted the Applicant Data to the Department, Goldman requested that it be excepted from public disclosure pursuant to Public Officers Law § 87(2)(d). That statute provides an exception from disclosure for records or portions thereof that “are trade secrets or are submitted to an agency by a commercial enterprise or derived from information obtained from a commercial enterprise and which if disclosed would cause substantial injury to the competitive position of the subject enterprise.”

Thereafter, the Department requested that Goldman submit a Statement of Necessity (“Statement”) pursuant to Public Officers Law § 89(5). Goldman submitted its Statement requesting that the Department continue to exempt the Applicant Data from public disclosure. In the Statement, Goldman asserts that the Applicant Data is protected under Public Officers Law § 87(2)(d), since disclosure would cause substantial injury to Goldman’s competitive position. Specifically, Goldman asserts that the Applicant Data contains its proprietary information, which includes its confidential credit underwriting policies and procedures, as well as voluminous records reflecting personal and financial data of applicants that Goldman Sachs deemed relevant to its confidential credit underwriting analyses. Accordingly, Goldman asserts that disclosure of its Applicant Data would cause it to sustain substantial competitive injury, since it would allow competitors to offer more favorable terms, conditions, rates, and benefits to customers without undertaking the significant time, effort and expense expended by Goldman to develop and test its underwriting models and formulas. Consequently, disclosing the Applicant Data would allow competitors to outcompete Goldman, thereby causing Goldman to sustain substantial competitive injury.

In Encore Coll. Bookstores, Inc. v. Auxiliary Serv. Corp., 87 N.Y.2d 410 (1995), the New York Court of Appeals provided the standard for determining whether records responsive to a FOIL request should be withheld from disclosure due to competitive injury concerns. In Encore, the Court held that determining whether an entity suffers competitive injury from public disclosure of its records within the meaning of Public Officers Law § 87(2)(d), “turns on the commercial value of the requested information to competitors and the cost of acquiring it through other means.” Id. at 420. The Court’s holding recognized that the trade secret/competitive injury exemption balances the need for openness in government versus the need “to protect businesses from the deleterious consequences of disclosing confidential commercial information, so as to further [New York] State’s economic development efforts and attract business to New York.” Id. Whether disclosing records, in whole or in part, would cause substantial competitive injury to an entity depends on, among other things, the nature of the information involved and the area of commerce in which the entity does business. See Committee on Open Government, Advisory Opinion 10664 (March 10, 1998).

Applying Encore, the Department agrees that disclosure of the Applicant Data would cause Goldman to suffer substantial competitive injury. A competitor who submits a FOIL request for the Applicant Data would certainly receive a windfall, since the competitor would have paid only a minimal fee, if any, to obtain proprietary information that Goldman has expended substantial resources to develop. Additionally, disclosing the Applicant Data would create an unlevel playing field, as Goldman does not possess the proprietary underwriting policies and procedures or applicant demographic information of its competitors. Finally, releasing the Applicant Data would undermine Goldman’s business interests and result in competitive harm, since Goldman’s customers rely on its ability to preserve the confidentiality of its business transactions and customer relationships and new and existing customers alike would move their business to banks that are able to preserve confidentiality. See N.Y. State Elec. & Gas Corp. v. NYS Energy Planning Board, 221 A.D.2d 121, 125 (3rd Dep’t 1996) (holding that respondent would be unable to fairly compete for customers in the future if the respondent could not guarantee that customer confidentiality would not be violated). Therefore, disclosing the Applicant Data would unfairly benefit Goldman’s competitors, thereby causing Goldman to suffer substantial injury to its competitive position.

Based upon Goldman’s Statement, and pursuant to Public Officers Law § 87(2)(d), the Department determines that the Applicant Data will not be disclosed, since disclosure would cause substantial injury to Goldman’s competitive position.

Pursuant to Public Officers Law § 89(5)(c), the requester may appeal the **Trade Secret Determination within seven business days** by sending an email to FOIL.Appeals@dfs.ny.gov.

III. Final Determination

Furthermore, the Department is withholding the Applicant Data from disclosure pursuant to Public Officers Law §§ 87(2)(a) and (b). Specifically, Public Officers Law § 87(2)(a) exempts from disclosure records exempted by a state or federal statute. The applicable statutory provision here is N.Y. Banking Law (“Banking Law”) § 36(10). Banking Law § 36(10) states, in pertinent part, that reports of examinations and investigations and correspondence and memoranda concerning or arising out of such examinations and investigations are confidential and shall not be made public. The statute fosters open communication between the Department and its regulated institutions, a necessity for effective regulation of financial institutions, by ensuring that records transmitted to the Department in connection with its supervision of a financial institution are protected from disclosure. The Applicant Data were transmitted to the Department in connection with the Department’s investigation of Goldman Sachs’ underwriting of Apple-branded credit cards and are, therefore, required to be kept confidential and exempt from disclosure under Banking Law § 36(10). Release of such sensitive records may have a chilling effect on the willingness of regulated entities to share information and cooperate with supervisors to resolve issues, thus, confidentiality is critical to the Department’s ability to perform its regulatory mandate and purpose. Accordingly, the exemption set forth in Public Officers Law § 87(2)(a) and Banking Law § 36(10) covers the Applicant Data.

Public Officers Law § 87(2)(b) (the “Personal Privacy Exemption”) prohibits the disclosure of responsive records or portions thereof that “if disclosed would constitute an unwarranted invasion of personal privacy under the provisions of [Public Officers Law § 89(2)].” Determining whether disclosure would constitute an unwarranted invasion of personal privacy requires balancing the competing interests of public access and individual privacy. See Dobranski v. Houper, 154 A.D.2d 736 (3d Dep’t 1989).

Here, the Department is withholding the Applicant Data for containing information such as application identification numbers and dates of births, since the interest in keeping that information private outweighs the competing interest of providing public access to that information. Disclosing this information about private citizens would not serve any governmental purpose consistent with the intent of FOIL. See Goyer v. New York State Dep’t of Envtl. Conservation, 12 Misc.3d 261 (Sup. Ct. N.Y. County 2005).

For the above-mentioned reasons, the Department has denied the FOIL request. Pursuant to Public Officers Law § 89(4), the requester may appeal the **Final Determination within thirty days** by sending an email to FOIL.Appeals@dfs.ny.gov.

Very truly yours,

Norma Roper Webster
Senior Attorney & Record Access Officer

cc: **VIA EMAIL**

(Shoshanah.Bewlay@dos.ny.gov)
Shoshanah Bewlay, Executive Director
Committee on Open Government
One Commerce Plaza
99 Washington Avenue-Suite 650
Albany, NY 12231

(skingsle@cs.cmu.edu)
Ms. Sara Kingsley
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213