当公平性审计的可行性取决于数据分享意愿：考察用户对用于公平性监测的多方计算协议的接受度

CHANGYANG HE*†，中国深圳哈尔滨工业大学和德国马克斯·普朗克安全与隐私研究所（MPI-SP）
PARNIAN JAHANGIRIRAD*，德国马克斯·普朗克安全与隐私研究所（MPI-SP）和萨尔大学，德国
LIN KYI，德国马克斯·普朗克安全与隐私研究所（MPI-SP）ASIA J. BIEGA‡，德国马克斯·普朗克安全与隐私研究所（MPI-SP）

arXiv:2602.01846v1 [cs.HC] 2 Feb 2026

公平性监测对于发现算法偏见至关重要，也是欧盟《人工智能法案》的要求。由于此类监测需要敏感用户数据（例如族裔），《人工智能法案》仅允许在采取严格隐私保护措施（如多方计算（MPC））并遵守《通用数据保护条例》的情况下处理这些数据。然而，此类安全监测协议的有效性最终取决于人们分享数据的意愿。目前，人们对不同的多方计算协议设计如何影响用户接受度知之甚少。为此，我们在欧洲对833名参与者开展了一项在线调查，考察用户对用于公平性监测的各种多方计算协议设计的接受度。研究发现表明，用户在直接评估中优先考虑风险相关属性（例如隐私保护机制），而在模拟选择中则优先考虑收益相关属性（例如公平性目标）；用户的公平与隐私取向也会影响其接受度。我们据此提出了相关建议，探讨如何部署和传达隐私保护协议，以促进知情同意并符合用户期望。

CCS 概念：• **安全性与隐私 → 安全与隐私的人文和社会层面**。

A附加关键词与短语：数据共享、多方计算、公平性监测、用户接受度

ACM 参考格式：
CHANGYANG HE、PARNIAN JAHANGIRIRAD、LIN KYI 和 ASIA J. BIEGA。2021。数据分享意愿决定公平性审计的可行性：考察用于公平性监测的多方计算协议的用户接受度。《*ACM*期刊》1，1，文章1（2021年1月），共34页。
https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

*这些作者对本研究的贡献相同。
†这项工作主要是在作者于马克斯·普朗克安全与隐私研究所（MPI-SP）任职期间开展的。
‡通讯作者。

作者联系信息：Changyang He，哈尔滨工业大学（深圳），中国深圳，以及马克斯·普朗克安全与隐私研究所（MPI-SP），德国波鸿，hechangy@hit.edu.cn；Parnian Jahangirirad，马克斯·普朗克安全与隐私研究所（MPI-SP），德国波鸿，以及萨尔大学，德国萨尔布吕肯，parnian.jahangirirad@mpi-sp.org；Lin Kyi，马克斯·普朗克安全与隐私研究所（MPI-SP），德国波鸿，lin.kyi@mpi-sp.org；Asia J. Biega，马克斯·普朗克安全与隐私研究所（MPI-SP），德国波鸿，asia.biega@mpi-sp.org。

特此授权以个人或课堂使用为目的，免费制作本作品的全部或部分数字版或纸质副本，但前提是不得为营利或商业利益制作或分发副本，且副本首页必须附有本声明及完整引文。本作品中由非 ACM 所有的组件，其版权必须得到尊重。在注明出处的情况下，可以进行摘要。除此之外，如需复制、再版、发布到服务器或重新分发至列表，必须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 联系以申请许可。© 2021 ACM。稿件已提交至 ACM

1 引言

公平性监测对于确保算法问责至关重要，尤其是在技术法规要求高风险应用不得产生歧视性结果的情况下。例如，欧盟《人工智能法案》将就业相关系统指定为高风险系统，并要求提供者对其进行偏见和歧视审查 [29]。然而，这项要求带来了一个根本性挑战：有效衡量公平性需要获取族裔或性取向等敏感用户数据 [34, 73, 84], ，但隐私法规（如欧盟《通用数据保护条例》（GDPR））对这些信息提供了严格保护 [101]。《通用数据保护条例》将此类信息指定为特殊类别数据，原则上禁止处理，尤其是在未获得有效同意的情况下 [54]。在招聘以及许多其他高风险应用中，有观点认为，由于求职者处于弱势地位，且求职者与雇主之间存在固有的权力差异，可能无法获得有效（自由给予的）同意 [17, 56, 77]。

为了解决这一矛盾，《人工智能法案》允许在特定条件下为公平性审计处理敏感数据（第 10 条第 5 款），其中包括部署“最先进的安全与隐私保护措施” [99]。人们提出了多种隐私保护方案，例如假名化、匿名化、差分隐私或多方计算（MPC），认为这些方案在技术上可行，可作为合规监测方法的基础，使得无需披露敏感数据即可计算公平性指标 [46, 51]。在这些方案中，多方计算尤其是一个很有前景的方向，因为它无需注入噪声即可保持数据保真度；这在公平性监测场景中尤为重要，因为此类场景下的测量必须应对有限的样本量 [114]。

然而，公平性测量是否可行，最终取决于最终用户是否愿意与服务提供商或审计人员分享其敏感数据。即使在技术上可以将隐私保护方法应用于公平性监测，此类安全监测框架的有效性仍取决于最终用户的接受和参与 [60, 76], ，而他们通常并非隐私保护计算方面的专业人士。然而，如何选择这些协议的具体部署设计，并以符合用户预期的方式与其沟通，这一挑战仍鲜少得到探索。近期的人机交互研究强调，如果缺乏能够促进信任、透明度和有意义的同意的以用户为中心的方法，隐私保护方案往往难以奏效 [25, 28, 39, 63, 111, 112]。这种以用户为中心的视角在公平性监测中尤为关键，因为建立不同用户群体之间的信任，对于确保获得具有代表性的数据集、从而得出有意义的结果至关重要 [35]。

为填补这一空白，我们考察用户对用于公平性监测的多种MPC协议设计的接受度。运用隐私计算理论 [24], ，我们将设计属性分为收益相关属性（公平性目标 和金钱激励）、风险相关属性（隐私保护机制 和 数据存储），以及同时与感知收益与风险相关的属性（收集的信息类型、监测主体 和 数据使用）。我们首先探究用户在决定是否参与公平性监测时如何进行隐私与收益权衡。随后，我们考察用户的公平性取向、隐私取向以及个人情境和人口统计特征作为隐私权衡 [45, 90, 109], 中的关键要素如何与协议接受度相关。我们提出以下研究问题：

- **RQ1**：用户对用于公平性监测的不同多方计算协议设计的优先考量与偏好是什么?
- **RQ2**：用户的公平与隐私取向及情境如何与用于公平性监测的多方计算协议的用户接受度相关?

为回答这些研究问题，我们对受《通用数据保护条例》和《人工智能法案》管辖的欧洲国家的 833 名在职求职者开展了在线调查。在算法招聘的公平性监测场景中，求职者代表着潜在的数据共享者。稿件已提交至 ACM

；这是一种具有代表性的数据共享情境，需要使用敏感数据来检测算法偏见。

针对 RQ1，我们通过直接属性排序测量用户的陈述的属性重要性，并通过基于情境的联合分析，针对不同设计属性测量用户的显现的属性重要性。我们的研究发现表明，在直接属性排序中，用户优先关注与隐私风险相关的属性，例如收集的信息类型和隐私保护机制；但在基于情境的研究中，他们更关注收益相关属性，例如公平性目标和金钱激励。用户表现出若干符合预期的隐私偏好，例如，与仅采用匿名化相比，他们更倾向于增加分布式数据存储；他们也更信任研究中心，而非商业公司。然而，更广泛的数据请求和扩大的数据使用却出人意料地提高了用户的数据分享意愿。我们推测，这是因为用户认为这些数据有助于实现系统的公平性目标。

针对 RQ2，我们通过回归分析发现，用户的公平与隐私取向塑造了其协议偏好：此前对数据共享持更开放态度的用户更强调公平性目标，而注重隐私的用户（即采取积极隐私保护措施且高度重视数据保护透明度的用户）较少关注公平性目标或金钱激励，而更多关注风险相关属性。基于这些研究发现，我们讨论了在《通用数据保护条例》和《人工智能法案》框架下，如何设计和传达隐私保护型公平性监测协议的相关启示。

本研究对 HCI 作出以下贡献：(1) 以人为中心地研究用户对 MPC 公平性监测协议的接受度，揭示不同设计属性之间细致的收益-隐私权衡；(2) 分析用户的取向与情境如何影响协议设计中的隐私权衡；(3) 为部署隐私保护协议及其沟通提供实践启示，以确保实现具有实质意义的知情同意。

2 相关工作

2.1 隐私保护的公平性监测

随着招聘和刑事司法等高风险领域越来越多地采用算法决策，人们日益担忧数据驱动的算法可能强化并放大历史与社会偏见 [62, 73]。对此，研究人员和从业者投入了大量精力来提升算法公平性，开发了偏见缓解 [38, 55] 和公平感知学习 [34, 66]等技术。尽管算法公平性取得了进展，但现实场景部署中的算法经常面临与训练环境显著不同的数据和运行条件（例如数据漂移 [27] 和概念漂移 [72]），从而限制了公平算法的稳健性和可解释性。因此，近期研究强调了公平性监测作为持续跟踪算法公平性的过程的重要性，即在部署后阶段进行监测 [8, 13, 34]。这与欧盟的《人工智能法案》 [98], 一致，该法案要求开展“上市后监测”，以确保高风险人工智能系统在其整个生命周期内始终符合监管要求。

与训练环境不同，训练环境在很大程度上假定可以获取敏感属性，因为公开数据集可供使用；而在现实环境中，部署后阶段的公平性监测不可避免地面临隐私约束 [22, 53, 96, 100, 114]。例如，对敏感数据的担忧给纽约市针对算法偏见的审计带来了重大挑战 [49]。因此，保护隐私的公平性监测成为一个关键议题。例如，在欧盟，为确保对高风险人工智能系统进行偏差检测和纠正，在严格必要的情况下，此类系统的提供者可根据《人工智能法案》第10条第5款例外处理特殊类别个人数据 [99]。但是，此类处理必须“遵循适当的保障措施，以保护基本权利

以及自然人的自由”。因此，算法公平性与隐私领域的研究人员提出了多种保护隐私的公平性监测协议 [22, 57, 114]。以下概述了几个值得注意的例子。

差分隐私（差分隐私）。差分隐私为限制个人数据的披露风险提供了严格的数学形式化保证 [22]。在公平性监测中，可以通过向聚合统计数据（例如混淆矩阵或群体特定公平性指标）注入经过精心校准的噪声来应用差分隐私。这种方法能够在防止原始数据或精确模型输出泄露的同时，对公平性进行保护隐私的评估 [22, 58, 114]。尽管如此，添加噪声不可避免地会降低数据粒度，并可能损害公平性评估的可靠性。Zaccour 等人发现，只有当样本量（n）较大且隐私预算（ε）并非极端时，受差分隐私保护的聚合数据才能保持公平性指标的准确性；例如，$\varepsilon \approx 0.5$ 且 $n = 1{,}000$ 时，或 $\varepsilon \approx 0.05$ 且 $n > 5{,}000$ 时，结果较为可靠；样本量较小 n 或隐私预算非常低 ε 会导致估计结果不可靠 [114]。

合成数据集。合成数据集通过统计方式再现现实世界数据模式，为公平性监测提供了一种替代方法，使分析无需依赖实际用户数据 [78, 96]。然而，其可靠性仍是一个关键问题，因为合成数据与实际用户数据之间不可避免的差异可能会损害公平性评估的有效性 [114]。合成数据也带来了一些伦理挑战，尤其涉及同意和“多样性漂洗” [107]。在公平性领域，合成数据被视为一种增加数据集多样性的方法，但这类合成数据仍植根于原始数据集中的偏见，并可能导致偏差 [107]。

多方计算（*MPC*）。MPC通过使多个参与方能够联合计算公平性指标，同时不向任何参与方透露原始敏感数据，为隐私保护型公平性监测提供了另一种有前景的方法 [31, 51, 83]。例如，波士顿女性劳动力委员会运用MPC衡量不同性别和种族群体之间的工资差距，无需任何参与方透露私人信息 [19, 65, 83]。一种常见的实现方式是安全两方计算，即由可信第三方与模型所有者合作，使用双方的加密数据计算公平性指标 [26, 51, 80]。

我们的贡献。与依赖噪声注入或分布近似的差分隐私和合成数据集不同，MPC能够直接基于加密数据计算公平性指标，在提供强有力隐私保障的同时，保持更高的保真度。然而，这一优势也带来了实际挑战。例如，如何确保个体能够接受涉及可信第三方的分布式数据存储，以及应由谁担任这一第三方，都是尚待解决的问题。为解决这些问题，我们采用以人为中心的方法，研究用户在不同设计条件下对MPC协议的接受度，旨在为协议设计提供参考，使其能够兼顾技术严谨性以及用户对公平性和隐私的期望。

2.2 隐私约束下的数据共享用户接受度

数据捐赠是指个体为公共利益自愿贡献个人数据，现已成为社会公益项目收集数据的重要方式 [47, 102], ，包括开发公平算法 [35]。例如，人机交互领域的研究人员采用数据捐赠来研究用户对短视频的参与度 [115], 、检测青少年网络风险 [86], ，并支持创伤知情设计 [116]。

然而，尽管用户共享数据的隐私保护协议不断取得进展 [39, 51],，现有的数据捐赠工作往往强调数据共享的开放性，却未能充分告知志愿者其数据将如何被使用 [47] 或受到保护 [16, 39]。了解用户对隐私保护协议下数据共享的接受度，不仅有助于实现数据捐赠的目标。稿件已提交至 ACM

也有助于实现数据捐赠的透明度 和有意义的同意 ， 同时还能为协议设计 提供宝贵见解，使其体现用户的期望 [25, 39, 63, 111]。

为实现这一目标，研究人员广泛探讨了可能影响隐私约束下的数据共享用户接受度的人因和技术因素 [39, 61, 87, 106]。一项研究方向考察了激励 对数据共享的影响，包括金钱激励 [92], 个人信息收益 [14], 以及对社会公益的贡献 [94]。例如，Bietz 等人发现，提供基于共享数据生成的汇总报告能够促进参与医疗网络调查 [14]。数据类型作为与隐私风险内在相关的因素，也显著影响了用户的数据分享意愿（例如，社交媒体帖子与搜索数据 [79], 、照片和视频与敏感性较低的形式 [61], 、GPS 定位与机密性较低的智能手机使用数据 [106]）。近期研究也开始关注隐私保护机制 及相应的隐私沟通在塑造用户数据共享决策中的作用 [25, 39, 76, 111]。例如，Franzen 等人研究了隐私沟通如何影响数据共享，并强调了其在帮助用户了解隐私保护机制和潜在风险方面的重要意义 [39]。Kacsmar 等人发现，尽管最终用户难以理解多方计算等隐私计算的抽象定义，但他们认为隐私计算的具体应用场景具有启发性且切实可行，这提高了他们对数据共享的接受度；然而，数据共享和分析的目的仍然是决定其态度的首要因素 [60]。最后，研究人员密切关注人的特征 （例如，数字素养 [7],、隐私顾虑 [79],和心理特征 [61, 92]）如何影响数据共享决策。基于这些考量，我们将在第 4 节详细说明如何对可能影响用户接受度的多方计算协议设计属性进行情境化处理。

我们的贡献。 与其他数据捐赠场景相比，公平性监测依赖敏感属性（例如性别、种族和其他人口统计因素）来评估和缓解算法偏见。敏感属性的使用形成了一种情境，用户可能需要权衡为促进算法公平性作出贡献所带来的益处与潜在隐私风险。因此，了解用户对协议设计的偏好，对于开发既能实现公平性监测又能保护隐私的用户信任协议至关重要。本研究旨在填补这一空白，并提出关于隐私保护型公平性监测协议设计与沟通的实践启示。

3 研究定位：用于公平性监测的多方计算协议

公平性监测的现实场景部署发生在特定的算法应用情境中。本文将研究置于算法招聘领域；这一应用尤其容易受到历史和社会偏见所导致的歧视性结果影响 [34, 50] ，因此根据《人工智能法案》被列为高风险应用。

本文假设采用多方计算协议作为安全与隐私保护措施，以保护用于公平性监测处理的敏感数据。多方计算允许多方利用各自输入共同计算公平性指标，而无需任何一方披露其输入值，从而保障个人敏感属性的机密性 [46, 51]。以公平性监测多方计算协议为例，安全两方计算的流程包括以下步骤：

(1) 生成并存放双方组件：招聘流程结束后，候选人可以通过由可信第三方（TTP）管理的服务，自愿分享其敏感数据。随后，TTP 会生成双方组件。具体而言，设 $x_i^{(\text{sens})} \in \mathbb{R}$ 为用户 i的敏感属性（例如性别、种族）， $r_i \sim \mathrm{R}$ 为可信第三方（TTP）生成的随机秘密值，$c_{i,\ 1} = x_i^{(\text{sens})} + r_i$和 $c_{i,\ 2} = x_i^{(\text{sens})} - r_i$将作为秘密份额生成。随后，这两个秘密份额会被存放至 TTP 和

招聘公司独立处理；这样一来，每位用户的敏感属性会被秘密拆分为两个掩码组件，并分发给不同的参与方，任何一方都无法单独推断出 $x_i^{(\text{sens})}$。

（2）**用于公平性计算的安全多方计算**：在公平性评估期间，敏感属性通过多方计算进行虚拟重构：

$$\langle x_i^{(\text{sens})} \rangle = \frac{c_{i,1} + c_{i,2}}{2}$$

值 $x_i^{(\text{sens})}$ 仍以秘密共享形式保存，并用于公平性指标的计算。例如，给定 $x_i^{(\text{sens})}$，群体 G（例如招聘或面试率）的平均结果为：

$$\mu(G) = \frac{\sum_{i=1}^{N} y_i \cdot \mathbb{1}\left[\langle x_i^{(\text{sens})} \rangle \in G\right]}{\sum_{i=1}^{N} \mathbb{1}\left[\langle x_i^{(\text{sens})} \rangle \in G\right]}.$$

它还可用于计算人口统计差异 [20]，以衡量算法公平性，其定义为不同群体之间平均群体结果的差异：

$$\delta = \mu(a) - \mu(b),$$

所有公平性计算均基于秘密共享输入 $x_i^{(\text{sens})}$ 进行。因此，在任何时候，任何一方都无法获知任何个人的敏感属性。

尽管用于公平性监测的多方计算协议在敏感数据的隐私保障与公平性评估的分析准确性之间取得了合理平衡，但它们在现实世界中的实现部署中也引入了若干复杂性 [26, 51, 65]，例如选择一个用户愿意放心托管其数据的可信第三方，向用户传达协议以促成有意义的同意，或在不同人口统计群体之间建立信任，以确保能够收集多样且具有代表性的敏感数据，从而实现准确的公平性监测。事实上，从理论到实际实施来设计隐私保护协议，并有效地向非专业人士传达这些协议，仍然是隐私保护技术 [1]在现实世界中得到采用的关键挑战。尽管近期研究探讨了专家观点 [1]以及最终用户态度 [60]对隐私保护计算（包括多方计算）可用性的看法，但对用户在不同具体协议设计下的接受度进行定量分析，或许能为现实世界中的协议的有效部署与传播提供必要的洞见。

为研究这一问题，我们首先在第 4 节推导出与用户风险与收益感知相关的公平性监测多方计算协议设计维度，并考察用户对不同协议设计的接受度和偏好（RQ1）。此外，我们还探究用户的公平与隐私取向及其所处情境如何影响其协议设计接受度（RQ2），以期为有效的协议沟通提供参考，确保不同群体都能在知情的情况下表示同意。

4 从协议设计到用户接受度：隐私权衡视角

受“5W1H”准则 [43, 59],的启发，我们讨论第3节所介绍的公平性监测多方计算协议的不同设计属性可能如何与用户对数据共享的认知相关，涵盖公平性目标 （Why）、收集的信息类型 （What）、数据存储 （Where）、监测主体 （Who）、金钱激励、数据使用 以及隐私保护机制 （How）等方面。随后，我们采用隐私计算理论（PCT） [24] 来构建分析框架，探讨这些设计属性如何影响个体在该协议中的数据分享意愿。PCT认为，个体在决定是否分享其信息时，会权衡感知收益 与感知风险 [24]。提交至 ACM 的稿件

隐私计算理论已被广泛用于理解数字环境下个体如何做出数据共享决策 [9, 67, 105]。基于 PCT，我们将公平性监测多方计算协议的设计属性分为收益相关属性（公平性目标 和金钱激励）、风险相关属性（隐私保护机制 和数据存储），以及同时与收益和风险相关的属性（收集的信息类型、监测主体 和数据使用）。

4.1 收益相关属性

4.1.1 公平性目标：数据捐赠有助于实现哪种公平性目标？ 数据捐赠使个体能够自愿分享个人数据，以支持集体目标，例如推动研究或促进社会公益 [14, 94]。因此，社会效益，即数据捐赠如何为公共利益做出贡献，是潜在数据捐赠者的一项内在激励 [2, 94]。大量研究表明，了解数据收集的目的与人们所表达的数据分享意愿相关（例如， [79, 93, 95]）。

在公平性监测的数据捐赠情境中，影响社会效益的核心设计属性是公平性目标的选择。算法公平性领域的研究人员提出了不同的公平性目标，用于评估和改进算法 [34]。不同的公平性目标可能对应着不同甚至相互冲突的价值假设 [41]。在各种公平性目标中，个体公平 和 群体公平 是应用最广泛的两种，尽管二者的目标往往相互冲突 [21, 73]。具体而言，个体公平 旨在让相似的个体获得相似的决策，而群体公平 强调平等对待不同群体 [41]。不同的公平性目标决定了评估和改进算法的方向，并影响潜在数据捐赠者对其价值和效用的感知。因此，我们研究不同公平性目标如何影响个体对公平性监测协议的接受度。

值得注意的是，向非专业人士解释算法公平性颇具挑战 [91, 113]。我们不着重介绍数学定义，而是阐明公平性目标背后的直观含义，帮助用户理解优化目标。具体而言，我们分别用“以相似方式对待不同人口统计群体”和“以相似方式对待相似个体”来概括群体公平性和个体公平，并将“选择最合格的候选人”作为用于比较的基础性能价值。纳入更细致的公平性定义，拓展了此前关于公众对公平性目标认知的研究（例如，二元准确率—公平性权衡 [75]）。尽管这些目标在实际操作中可能有所重叠（例如，收集的数据可用于群体和个体公平性监测），但考察用户如何比较这些目标，有助于深入了解他们的公平性预期。不过，由于用户可能会以不同方式理解公平性目标之间的兼容性，因此应谨慎看待相应的研究发现——感知重要性较低并不意味着实际价值较低。

4.1.2 金钱激励：如何补偿数据捐赠者？随着各组织寻求鼓励更多人参与，经济激励已成为数据共享领域备受关注的话题 [37, 61]。尽管金钱激励似乎是提高参与度的显而易见手段，但相关文献指出了若干值得关注的问题 [37, 92]。例如，Fast等人发现，金钱激励能在短期内促进数据捐赠，但由于参与者感受到机会成本且感知收益有限，这种效果未能持续 [37]。Silber 等人指出，高额激励可能引起受访者的怀疑，使他们认为自己的数据价值很高，最终可能降低其参与意愿 [92]。此外，还有人担心经济激励可能会不当地影响或胁迫个体，并损害知情同意 [3]。与此一致，《通用数据保护条例》要求同意必须在不受胁迫的情况下自由作出 [32, 54]。

在本研究中，我们纳入金钱激励，以考察参与者如何权衡其相对于其他协议设计属性的重要性。然而，对于金钱激励情境下的数据分享意愿，应结合伦理和法律考量谨慎解读 [11]。

4.2 风险相关属性

4.2.1 隐私保护机制：数据如何受到保护？研究人员越来越多地关注支持数据捐赠的隐私保护框架和技术 [4, 15, 51]。然而，由于隐私保护机制通常被认为过于复杂，难以确保参与者充分了解相关信息，因此只有少数研究考察了这些机制对个体数据共享决策的影响 [25, 39, 111]。例如，Franzen 等人的近期研究表明，通过结合图形化风险可视化和交互式风险探索，帮助用户理解隐私保护机制（在其研究中为差分隐私）能够促进用户就数据共享作出知情决策 [39]。Xiong等人发现，与展示定义或流程相比，展示解释其影响的描述有助于参与者更好地理解差分隐私 [111]。

尽管多方计算提供了技术上强大的隐私保护机制，但用户如何信任并接受多方计算来保护共享数据仍不明确，因为多方计算涉及更多利益相关方和更复杂的数据处理。在本研究中，我们探究多方计算的隐私保护机制如何影响用户对公平性监测协议的接受度。借鉴先前研究的见解 [39, 111], ，我们选择不呈现多方计算的技术定义。相反，我们使用可视化向研究参与者传达多方计算实现隐私保护的直观基础（分布式存储 +加密计算），如附录B所示。这有助于与问卷调查研究中的基线条件进行比较，即匿名化和加密。

4.2.2 数据存储：加密数据存储在何处？ 近期研究表明，存储数据的实体会影响参与者分享敏感数据的意愿 [52, 87]。一般而言，对数据存储位置的感知安全性及其声誉在塑造信任方面起着关键作用。例如，德国的一项问卷调查研究发现，参与者更倾向于将捐赠数据存储在全国性集中式数据库中，而非长期存储在研究机构 [87]。美国的一项在线实验也表明，人们更希望自己的数据由学术机构或政府机构管理，而非私人实体 [52]。

多方计算协议假定存在一个独立于运营求职申请系统的服务提供商的单独且独立的实体，作为可信第三方（可信第三方），以支持敏感数据的分布式存储以及加密的公平性计算。因此，它既需要具备必要基础设施的技术能力，也需要承诺保护数据捐赠者的隐私 [5, 51]。确保用户对可信第三方的信任，对于多方计算协议的现实世界部署至关重要。

因此，我们考察用户如何将数据存储位置与其他协议设计属性进行优先级排序，以及他们如何看待、信任并接受不同的可信第三方作为数据存储地点，包括研究中心、非政府非营利组织和专业商业公司。

4.3 与感知收益与风险均相关的属性

4.3.1 收集的信息类型：收集了哪些数据？所请求数据的类型和范围会影响个体对隐私风险和数据捐赠价值的认知，进而影响数据捐赠决策 [18, 61, 92, 106]。具体而言，个体对所收集数据的感知可识别性和敏感性，是影响其数据分享意愿的重要因素 [61]。例如，Wenz等人发现，与加速度计数据或问卷数据相比，参与者分享带有地理位置信息的智能手机GPS数据的意愿较低。提交给 ACM 的稿件

同样，Silber 等人发现，健康应用数据的总体分享率（6.1%）远低于 Twitter（24%）等社交媒体数据 [92]。本研究聚焦于算法招聘中的公平性监测，并将收集的数据类型划分为三种可能的条件：

(1) 人口统计信息 （例如年龄和性别），这是衡量输入公平性所需的最少数据 [85], ，反映候选人群体的构成是否公平且具有代表性；(2) 人口统计信息 +招聘决策 （例如面试邀请或工作邀约），能够支持对结果公平性进行更丰富的监测，例如人口统计均等（招聘系统结果中不同人口统计群体获得的结果具有 [21]）；以及(3) 人口统计信息 + 招聘决策 +资历 （例如教育背景和工作经验），能够支持对输出公平性（例如排名结果中不同群体获得的曝光机会均等 [34]）和结果公平性（例如不同群体中资历相当的个体享有机会均等 [21, 34]）进行更复杂的监测，而这需要模型输入。

这三种条件支持复杂程度各异的公平性监测任务，并产生丰富程度不同的结果；每种结果都对应着不同程度的隐私风险和感知有用性。在本研究中，我们考察所收集的数据类型的变化如何基于这三种条件影响参与者对公平性监测协议的接受度。

4.3.2 监测主体：谁在收集用于公平性监测的数据？管理公平性监测数据捐赠项目的利益相关方也会显著影响数据分享意愿。就感知收益而言，个体对不同组织为社会福祉收集数据的信任程度各不相同 [71]。就感知风险而言，与私营企业相比，个体对学术机构等声誉良好的组织较少担心隐私侵犯 [52, 71]。

在我们的研究中，我们假设了由不同监测主体负责的两种数据捐赠情景：（1）由审计机构和监管机构（外部利益相关方）组织数据捐赠；（2）由开发招聘算法的公司（内部利益相关方）组织数据捐赠。通过调查，我们评估负责管理数据捐赠项目的不同利益相关方如何影响参与者对该协议的接受度。

4.3.3 数据使用：数据用于何处？根据欧盟《人工智能法案》第10条第5款， [97], 人工智能提供者“为确保对高风险人工智能系统进行偏差检测和纠正，可在特殊情况下处理特殊类别个人数据，但须采取适当保障措施，以保护自然人的基本权利和自由”。这一例外仅适用于“训练、验证和测试数据集” [97]。尽管此类法律法规界定了数据使用的条件和范围，但公众如何接受不同程度的数据用于公平性监测，仍是一个值得关注的问题。事实上，数据使用可能影响人们对数据捐赠的感知收益与风险。当用户数据仅用于公平性评估，以检测偏见或歧视行为时，由于不会直接影响算法，隐私风险相对较低。然而，要提升算法公平性，还需要采取额外的公平性干预措施，例如使用合成数据集，这可能会降低用户感知到的效用 [23, 70]。另一方面，除了用于评估之外，使用用户共享数据训练公平算法，虽能更直接地影响算法，却会加剧安全性和隐私顾虑，包括训练数据推断攻击可能导致敏感信息暴露 [74]。

因此，我们在调查中提出了两种情境：（1）数据仅用于模型评估；（2）数据用于模型评估和开发。我们考察不同的数据使用方式如何影响参与者在该协议下的数据分享意愿。

5 方法

在本节中，我们介绍用于考察 RQ1 和 RQ2 的调查工具。针对 RQ1，我们在第 5.1 节开展（a）基于情境的联合分析以及（b）直接属性排序任务，以评估用户对协议属性感知的重要性，并结合属性级评分来衡量用户对不同属性水平的偏好。针对 RQ2，我们在第 5.2 节运用回归分析，以刻画用户的公平与隐私取向及情境如何与其对协议设计的接受程度相关。最后，我们在第 5.3 节概述最终研究设计，并详述参与者招募与调查流程。招聘及研究程序已获得我们机构审查委员会（IRB）的批准。

5.1 RQ1：理解用户对用于公平性监测的多方计算协议不同设计的优先事项与偏好

为探究用户对 MPC 公平性监测协议各种设计的优先事项与偏好（RQ1），我们首先通过两种互补方法考察用户如何权衡不同的设计属性：基于情境的联合分析用以衡量揭示的属性重要性，以及直接属性排序用以衡量陈述的属性重要性。此外，我们纳入属性级评分问题，以更细致地评估用户对特定属性水平的偏好。

5.1.1 通过联合分析测量揭示的属性重要性。 我们首先采用联合分析，研究公平性监测多方计算协议的不同设计属性如何影响用户对数据共享的决策 [48]。联合分析是一种常用的统计技术，用于考察个体如何通过权衡各种属性来做出复杂决策。该方法最广泛应用于营销领域，并日益受到人机交互和可用隐私研究的重视，这些研究探索数字应用的设计空间 [6, 64,82]。我们通过一系列针对不同协议设计的假设性情境选择，测量用户如何权衡不同设计属性（如表1所示）；这些属性的选取依据第4节中的讨论。该方法有助于大规模比较评估不同设计方案，使我们能够深入了解用户如何应对数据共享中固有的权衡，尤其是在促进公平性与保护隐私之间的权衡。

每位参与者都完成了五项联合分析任务，并从中选出自己偏好的协议设计。每项任务包含两个涵盖七项属性的设计选项，以及一个“以上皆非”选项，以减轻强制选择偏差 [48]。图1展示了一项任务示例。属性及其水平的组合共产生864种可能的协议设计版本（$3 \times 3 \times 2 \times 4 \times 3 \times 2 \times 2$）。为控制复杂度并减少向参与者展示的版本数量，我们使用Sawtooth Software[1] 生成正交设计，最终得到300种独特的协议设计版本 [6]。

基于参与者的选择，我们使用 Sawtooth Software 实现的层级贝叶斯（*HB*）模型估算了个体层面的部分效用。该模型假设受访者的选择遵循多项逻辑回归模型，即选择某一选项的概率取决于该选项相对于其他选项的效用。效用估计的详细说明见附录 A。随后，我们推导出属性重要性

[1]https://sawtoothsoftware.com/ 稿件已提交至 ACM

表 1. 基 r于选择的联合分析调查中呈现的属性 。

类别	属性	描述	可能的取值
收益相关属性	公平性目标	哪种公平性目标数据捐赠贡献于?	（1）选择最合格的候选人；（2）平等对待不同的人口统计群体；（3）平等对待相似的个体
	金钱激励	数据捐赠者如何获得补偿?	(1) 无金钱补偿；(2) 1美元礼品卡；(3) 5美元礼品卡
风险相关属性	隐私保护机制	数据如何受到保护?	(1) 匿名化并加密个人数据；(2) 匿名化并加密个人数据，然后分发加密数据，使任何单一方都无法完全访问敏感个人数据。
	数据存储	已加密内容存放在哪里数据存储在何处?	（1）政府；（2）研究中心；（3）非政府非营利组织；（4）专业化商业公司
属性相关对象 Both	收集的信息类型	收集了哪些数据?	（1）人口统计属性；（2）人口统计属性与招聘决策；（3）人口统计属性、资历与招聘决策
	监测 利益相-关者	谁在收集数据用于公平性监测?	（1）审计机构和监管机构；（2）开发自动化招聘系统的公司
	数据使用	数据用于什么？	（1）评估自动化招聘系统；（2）设开发和评估自动化招聘系统

通过计算每个属性内部分效用的相对范围，并在所有属性之间进行归一化，使其总和为 100%。

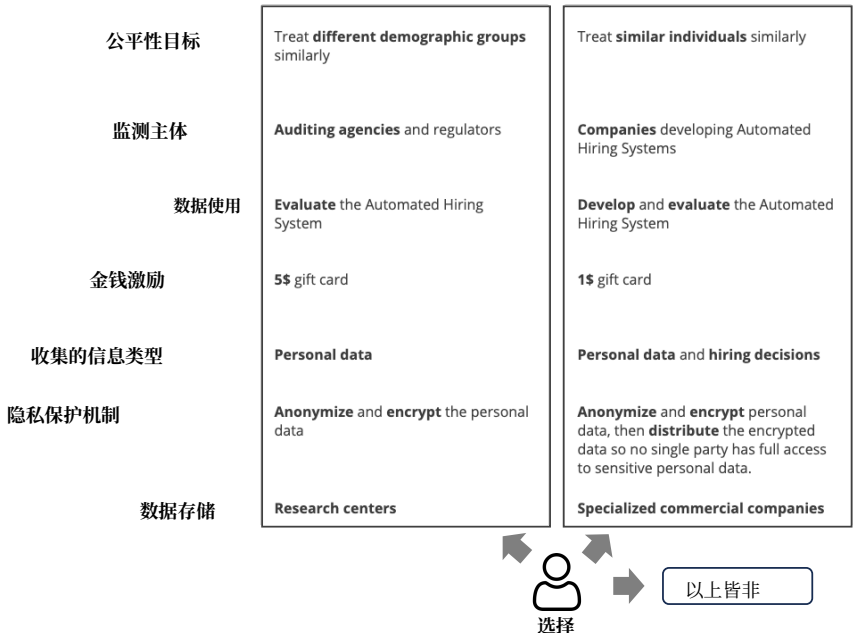
5.1.2 通过直接属性排序衡量陈述属性重要性。除了从联合分析中推断出的揭示属性重要性外，我们还衡量了陈述属性重要性，使用户能够在属性排序任务中直接表明其设计优先级：调查直接呈现七个属性，参与者按照重要性从高到低对属性进行排序。为估计重要性，我们采用了反向计分，即数值越高表示感知到的重要性越高。这种直接排序方式补充了联合分析任务 [30]；排序能够简单、直观地表达感知到的重要性，而基于选择的联合分析则促使参与者进行类似现实世界决策的真实权衡。

5.1.3 通过属性水平评分衡量属性设计偏好。在考察用户如何权衡不同设计属性之后，我们进一步研究用户对每个属性内具体属性水平的偏好。具体而言，我们针对七个属性的每个属性水平设计了 7 点李克特量表问题，使参与者能够在 -3（最不可接受）到 +3 （最可接受）的范围内评价其对每个属性水平的接受程度。收集数据后，我们使用 Welch 方差分析检验 [69], 比较每个属性内不同属性水平之间的差异，因为用户的评分近似呈正态分布，但存在方差不等。

5.2 RQ2：探讨用户的公平性与隐私取向及其情境如何与协议设计接受度相关

在了解用户对多方计算协议设计的优先事项与偏好后，我们进一步通过

图 1. 用于考察用户在协议设计属性之间权衡的联合分析任务示例。受访者在两种 MPC 公平性监测协议设计之间进行选择，或选择“以上皆非”。通过联合分析对其选择进行分析，以揭示属性重要性。



回归分析探究用户的公平与隐私取向及情境与其对协议设计的接受度之间的关联。这项研究有助于设计更具包容性的协议，并支持定制化的协议沟通，从而促成有实质意义且充分知情的同意。

5.2.1 因素构建。以隐私权衡视角为指导，鉴于用户的取向及情境在隐私权衡中发挥着重要作用，我们将用户的公平与隐私取向及情境操作化为可能影响其接受用于公平性监测的多方计算协议的因素 [45, 90, 109]。具体而言，我们纳入隐私取向，以体现用户对隐私风险的态度；纳入公平取向，以体现用户对公平性监测的重视程度；并将人口统计特征、以往经历和系统素养作为影响公平与隐私感知的情境因素 [104, 118]。本研究涵盖以下因素：

- **隐私取向**：用户对隐私风险和控制措施的立场，通过以下指标衡量：总体数据捐赠意愿、对数据捐赠的感知风险、主动隐私保护措施、对在线隐私风险的理解、对在线数据使用的理解、对数据使用透明度的感知重要性，以及对数据保护透明度的感知重要性。

- **公平取向**：用户对公平性监测的重视程度，通过以下指标衡量：对公平性监测系统有效性的感知以及对算法招聘系统感知公平性的评价。
- **系统素养**：用户对相关系统的理解和效能感，包括对自动化招聘系统的理解以及对公平性监测系统的理解。
- **以往经历**：任何情境下的个人歧视经历、求职申请中的个人歧视经历、亲近关系中的感知歧视经历，以及以往的数据捐赠经历。

- **人口统计特征**：年龄、性别、族裔、教育背景和就业状况。

5.2.2 分析。 我们进行了线性回归分析，以探究用户的公平与隐私取向及情境与用户对协议的接受度之间的相关性（RQ2）。具体而言，我们针对七项协议设计属性分别进行了回归分析，并采用两种不同的结果指标：（1）通过联合分析任务的层次贝叶斯（HB）模型估算的揭示重要性；（2）从排序任务中获得的陈述重要性。共得到14个回归模型（7项属性 $\times 2$ 结果类型）。我们对数值特征进行了标准化，并使用独热编码表示对分类变量进行编码。我们采用L1正则化，通过将较不重要的系数收缩至零来筛选特征，从而减少过拟合并提升可解释性。

5.3 研究设计

本 节概述研究设计的具体细节，包括调查流程和参与者招募 nt.

5.3.1 调查流程。 基于第4节的理论基础，三位作者协作并反复完善调查问卷，并通过多轮讨论解决分歧。随后，我们邀请欧盟的八位参与者开展预研究，他们来自不同的文化背景（欧洲、北美和亚洲），对隐私增强技术和算法公平性的熟悉程度也各不相同。预研究的反馈有助于改进调查问卷，消除令人困惑的内容，并确保不同知识水平的参与者都能顺利完成调查。完整的调查问卷详见附录B。

背景知识介绍。 为帮助参与者理解本研究，调查首先介绍了关键背景概念，包括自动化招聘系统、此类系统中的公平性、公平性监测中数据捐赠的作用，以及利用多方计算保护敏感数据的基本理念。为减轻认知负担，我们采用了视觉示例（见附录B）来说明：（1）有偏训练数据如何导致自动化招聘系统出现公平性问题；（2）数据捐赠如何帮助发现此类问题；（3）如何结合TTP辅助的分布式存储与多方计算，在不泄露敏感信息的情况下开展公平性监测。为确保参与者充分理解背景材料，我们设置了三道二选一评估问题。未通过评估的参与者会被重新引导至介绍页面，并须正确回答全部三道问题后才能继续参与调查。

收集用户的公平与隐私取向及情境。 在向参与者提供必要的背景知识后，我们按照第5.2节所述，收集了用户的公平与隐私取向及情境信息，包括人口统计特征、以往经历、系统素养、公平取向和隐私取向。本节还包含一道注意力检查问题；回答错误的参与者将被排除在分析之外。

联合分析任务。 接下来，参与者完成了五项联合分析任务，以表明他们对协议设计的偏好，具体内容见第 5.1 节。根据预研究的反馈，参与者认为联合分析任务直观易懂，便于比较各项属性，因此我们在所有任务中保持属性顺序一致，以避免重复任务造成混淆；结果证实，属性顺序对揭示重要性没有显著影响。我们还将注意力检查安排在调查的其他部分，而非联合分析任务中，因为大多数属性都没有明显优于其他水平的优势水平（例如，更多的金钱激励可能会加剧感知风险；收集的数据较少可能会降低用户对数据捐赠的感知有用性）。因此，我们无法设计出一个明显优于其他选项的选项来进行注意力检查 [6]。

自我声明的属性重要性和水平接受度。 完成联合分析任务后，参与者被要求完成一项属性排序任务，以说明其陈述属性重要性，如第 5.1.2 节所示；并完成一项属性水平评分任务，以指出其偏好的设计选项，如第 5.1.3 节所示。

5.3.2 参与者招募。 在招募参与者之前，我们进行了功效分析，以确定所需的样本量。结果表明，假设每位参与者完成五项联合分析任务，至少需要 626 名参与者 [40, 89]。随后，我们通过 Prolific 平台招募参与者。[2] 符合条件的参与者须居住在实施《通用数据保护条例》的欧洲国家，且为当前求职者；这是多方计算协议在招聘场景中进行公平性监测所针对的群体。鉴于不同人口统计群体的公平性感知和隐私感知可能存在差异，我们分批招募参与者，并监测人口统计分布。由于最初招募的样本以白人为主，无法充分代表欧盟的人口统计特征，我们特意从非白人参与者中额外收集了 300 份答卷，以提升人口多样性。共有 833 名个体参与了本研究。调查预计耗时约 10 分钟，参与者可获得 2 欧元报酬。

剔除未通过注意力检查或在不到 5 分钟内完成调查的参与者后，最终样本由 795 名参与者构成。样本中男性略占多数：53.5% 的参与者认同为男性（$N = 425$），45.0% 认同为女性（$N = 358$），1.0% 认同为非二元性别（$N = 8$）。样本的种族分布如下：60.4% 为白人（$N = 480$），15.1% 为亚裔（$N = 120$），15.0% 为黑人（$N = 119$），7.9% 具有混合种族背景（$N = 63$），1.5% 来自其他群体（$N = 12$）。参与者中，年龄最大的群体为 25–34 岁，占样本的 39.9%（$N = 317$）；其次是 18–24 岁群体，占 21.9%（$N = 174$）；35–44 岁群体占 20.8%（$N = 165$）。45 岁及以上的受访者所占比例较小，其中 12.5% 为 45–54 岁（$N = 99$），4.2% 为 55–64 岁（$N = 33$），0.9% 为 65 岁及以上（$N = 7$）。总体而言，74.2% 的参与者接受过大学或更高程度的教育（$N = 590$），24.2% 完成了中等教育（$N = 192$），1.3% 接受过基础教育或未接受过教育（$N = 10$）。就就业状况而言，56.7% 的参与者处于就业状态（

$N = 450$），39.4% 处于失业状态（$N = 313$），其余用户未透露自身状态。用户在公平与隐私取向、以往经历和系统素养方面的均值与标准差见图2。

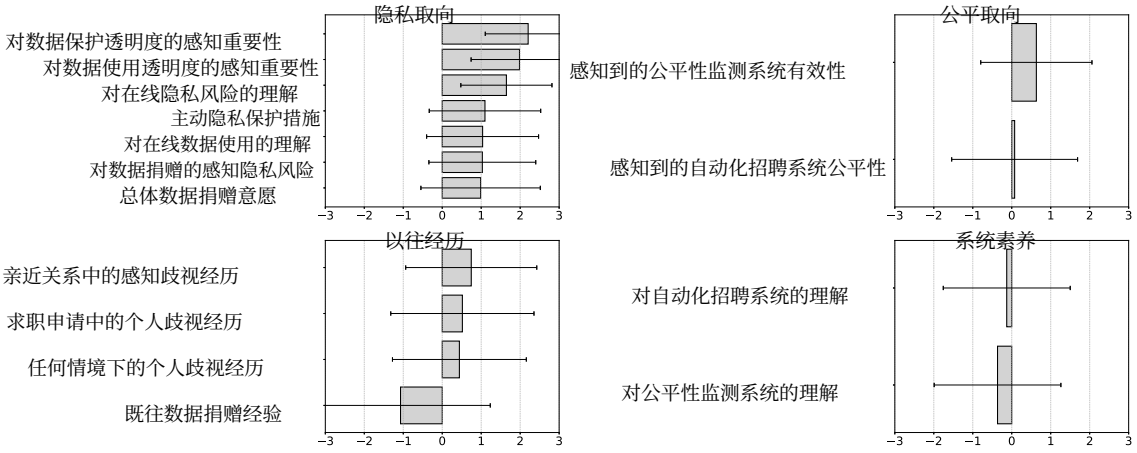


图 2. 用户公平与隐私取向、以往经历和系统素养的均值与标准差

[2]https://prolific.co/

6 研究发现

本节介绍用户对用于公平性监测的多方计算协议的接受度研究发现。我们将在第6.1节说明用户对不同设计属性的优先级和偏好（RQ1），并在第6.2节剖析用户的公平与隐私取向及情境与其对协议设计的接受度之间的关联（RQ2）。

6.1 RQ1：不同 MPC 公平性监测协议设计的用户接受度

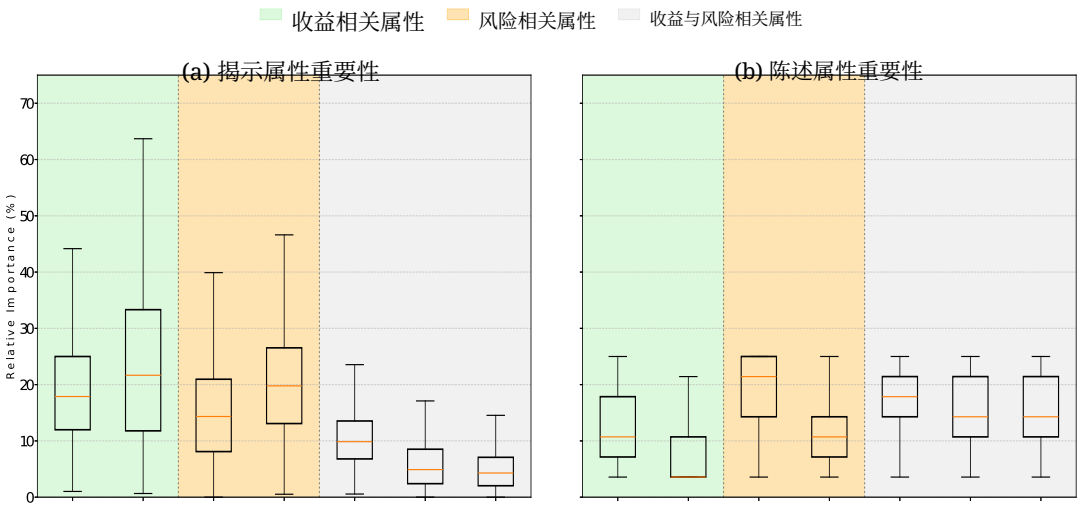


图3。属性联合分析揭示的重要性 MPC 公平性监测协议设计中的重要性，包括（a）基于选择型以及（b）基于属性排序的陈述重要性 d

6.1.1 MPC 公平性监控协议设计中的属性重要性。图 3 展示了 MPC 协议设计中各属性的重要性，包括：(a) 基于情境的联合分析得出的显现重要性，以及 (b) 基于直接属性排序得出的陈述重要性。

基于联合分析揭示的属性重要性：参与者在模拟决策情境中优先考虑利益相关因素。收益相关属性中的两项——金钱激励 和公平性目标——对用户选择偏好协议设计的决策有显著影响，重要性分别位列第一（均值 = 23.6%，标准差 = 14.4%）和第三（均值= 19.2%，标准差 = 9.8%）。在风险相关属性中，数据存储 被认为是第二重要的属性（均值 = 20.3%，标准差 = 10.0%），而隐私保护机制 的平均相对重要性为15.1%（标准差 = 9.2%）。涉及收益和隐私风险的三项属性——收集的信息类型、监测主体 和数据使用——其相对重要性分别为10.8%（标准差 = 5.8%）、6.0%（标准差 = 4.6%）和5.0%（标准差= 3.9%）。

基于直接排序任务陈述的属性重要性：参与者在直接排序中优先考虑收集的信息类型和隐私保护机制。直接排序所反映的用户陈述属性重要性，与用户在模拟联合分析任务中揭示的偏好存在显著差异。在排序任务中，优先级最高的属性是隐私保护机制 （均值 = 19.1%，标准差 = 5.4%）。收集的信息类型 在陈述重要性中位列第二（均值 = 17.6%，标准差 = 6.0%）。相比之下，尽管两项收益相关属性在模拟情境中对用户决策有显著影响，公平性目标 和金钱激励的重要性排名却处于较低水平，分别为12.0%（标准差 = 6.9%）和7.8%（标准差 = 6.4%）。

6.1.2 协议设计中不同属性水平的用户偏好。

总体协议设计偏好。表2展示了用户对协议设计中不同属性水平的偏好。对于公平性目标而言，一般算法性能（即选择资历最合格的候选人）优于群体公平（即以相似方式对待不同人口统计群体）和个体公平（即以相似方式对待相似个体），获得了最高的接受度得分（平均接受度得分 = 79.1%，标准差 = 20.8%）。对于金钱激励而言，更高的金钱补偿与更高的接受度得分相关。

我们的研究发现表明，参与者在为算法公平性作出贡献时，可能愿意考虑部分隐私让步。对于收集的信息类型而言，与仅收集基础人口统计数据相比，当协议收集所有相关信息（包括招聘决策、资历和人口统计特征）时，用户的接受度得分更高（平均接受度得分= 75.3%，标准差 = 24.6%）。同样，在数据使用方面，与仅评估自动化招聘系统相比，用户表现出更高的意愿，愿意为开发和评估该系统贡献其数据（平均接受度得分 =78.5%，标准差 = 19.2%）（平均接受度得分 = 72.0%，标准差 = 19.6%）。

总体而言，用户更信任无利益关系的机构，而非商业公司，来同时担任监测主体和数据存储的可信第三方（TTP）。具体而言，在担任数据存储的 TTP 方面，研究中心获得的接受度得分（平均接受度得分 = 76.9%，标准差 = 20.6%）高于非政府非营利组织（平均接受度得分 = 72.8%，标准差 = 23.3%）和专业商业公司（平均接受度得分 = 56.7%，标准差 = 29.6%）。此外，与开发自动化招聘系统的公司（平均接受度得分 = 63.5%，标准差 = 26.2%）相比，用户对由审计机构和监管机构担任监测主体 的接受度更高（平均接受度得分 = 73.6%，标准差 = 21.8%）。

最后，用户认为分布式存储——多方计算协议的关键隐私保护机制——比基础匿名化和加密更有效。尽管对个人数据进行匿名化和加密已获得较高的接受度得分（平均接受度得分 = 76.2%，标准差 = 19.5%），但加入分布式存储后，接受度进一步提高了约 10%（平均接受度得分 = 86.5%，标准差 = 18.7%）。

风险—收益权衡因人口统计群体而异。如表3所示，我们比较了不同人口统计群体对协议设计的接受度。在不同年龄组中，与年轻参与者相比，年长参与者更愿意在没有金钱补偿的情况下共享数据（平均接受度得分 = 46.4%）；而在提供5美元补偿的情况下，年轻参与者对数据共享的接受度更高（平均接受度得分 = 81.2%），年长参与者则为（平均接受度得分 = 73.7%）。与年轻参与者相比，年长参与者也更信任由审计机构和监管机构担任监测主体（平均接受度得分 = 77.2%），而年轻参与者对由开发自动化招聘系统的公司担任这一角色的接受度更高（平均接受度得分 = 64.7%），年长参与者则为（平均接受度得分 = 57.9%）。

表2. 协议设计中不同属性水平的用户偏好。“平均接受度得分”呈现参与者对每个属性水平的平均接受度得分，其中 100% 表示“最可接受”，0 表示“最不可接受”。所有属性的差异在 Welch 方差分析检验下均具有统计学显著性（*** $p < 0.001$）[69],该检验适用于我们近似正态分布且方差不等的设定。

属性水平	平均接受度（%）	标准差（%）
公平性目标* **		
选出最合格的候选人	**79.10**	20.80
平等对待不同的人口统计群体	73.77	25.08
平等对待相似的个体	68.76	24.09
金钱激励* **		
不提供金钱补偿	40.31	30.02
1美元礼品卡	58.93	24.26
5美元礼品卡	**79.90**	24.20
隐私保护机制* **		
匿名化和加密	76.21	19.48
匿名化、加密和分布式存储	**86.48**	18.72
数据存储* **		
专业商业公司	56.73	29.57
研究中心	**76.86**	20.60
非政府非营利组织	72.83	23.27
收集的信息类型* **		
人口统计数据	66.69	23.86
人口统计数据和招聘决策	70.23	21.39
人口统计数据、招聘决策和资历	**75.26**	24.59
监测主体* **		
审计机构和监管机构	**73.56**	21.77
开发自动化招聘系统的公司	63.52	26.24
数据使用* **		
评估自动化招聘系统	72.01	19.63
开发和评估自动化招聘系统	**78.51**	19.18

表3. 不同协议设计的用户接受度在人口统计群体之间的比较。我们用“His. Adv.”表示劳动力市场中历史上占优势的群体（年轻人/男性/白人求职者），用“His. Disadv.”表示与之相对的历史上处于弱势的群体。我们仅展示在 t 检验中且经 Benjamini-Hochberg 校正以控制假发现率后，群体间存在统计显著差异($p \leq 0.05$)的属性水平 [10]。*** $p < 0.001$，** $p < 0.01$，* $p \leq 0.05$

属性水平	平均接受度得分 历史上占优势的群体所占比例（%）	平均接受度得分（%）的他的。Dis-adv.
年龄：年轻人（<45）vs. 年长者（45+）		
金钱激励：无金钱补偿*	39.02	**46.40**
金钱激励：5美元礼品卡*	**81.20**	73.74
监测主体：审计机构和监管机构*	72.79	**77.22**
监测主体：开发自动化招聘系统的公司*	**64.71**	57.91
族裔：白人 vs. 非白人		
隐私保护机制：匿名化和加密***	73.61	**80.16**
数据存储：专业商业公司*	54.00	**60.90**
数据使用：开发和评估自动化招聘系统 *	76.91	**80.95**
监测主体：开发自动化招聘系统的公司*	61.49	**66.61**
性别：男性 vs. 其他性别		
金钱激励：5 美元礼品卡*	**81.96**	77.52

与白人参与者相比，非白人参与者在隐私约束下更愿意为算法公平性贡献自己的数据。具体而言，在隐私保护水平仅为基础程度（即匿名化和加密）时，非白人参与者（平均接受度得分 = 80.2%）比白人参与者（平均接受度得分 = 73.6%）更愿意分享自己的数据。此外，当商业公司（如表2所示，通常是信任度较低的实体）担任数据存储地点和监测主体时，非白人参与者对数据捐赠的接受度（平均接受度得分分别为 = 60.9 % 和 66.6%）高于白人参与者（平均接受度得分分别为 = 54.0% 和 61.5%）。最后，当数据用于开发和评估自动化招聘系统时——这一场景对算法的影响更为直接，但隐私泄露风险也更高——非白人参与者贡献数据的平均接受度得分（81.0%）显著高于白人参与者（76.9%）。

在性别方面，男性对金钱激励较高的协议表现出更高的接受度（平均接受度得分= 82.0%），高于其他参与者（平均接受度得分 = 77.5%）。此外，其他参与者比男性更重视群体公平（平均接受度得分 = 75.6% 对比平均接受度得分 = 72.2%），差异达到边缘显著水平（p = 0.052）。

主要发现

我们发现，在直接属性排序中，用户优先考虑与隐私风险相关的因素，即收集的信息类型 和隐私保护机制。然而，在模拟决策情境中，他们更关注利益相关因素，包括公平性目标 和金钱激励。从协议设计偏好来看，用户的隐私顾虑体现在多个方面，例如，相较于基础教育水平的匿名化和加密，他们更偏好分布式数据存储；在分布式数据存储中，他们也更信任研究中心而非商业公司担任可信第三方。令人意外的是，更广泛的数据请求和更充分的数据使用与更高的接受度相关，这可能是因为用户认为自己为算法公平性作出了有意义的贡献。这些风险—收益权衡也因人口统计群体而异。例如，在隐私条件较不利的情况下，非白人参与者比白人参与者显著更愿意为公平性监测分享数据。

6.2 RQ2：用户的公平性与隐私取向及情境与协议设计接受度之间的相关性

6.2.1 乐于共享数据的用户优先考虑公平性目标，而注重隐私的用户则较少重视公平性目标或金钱激励。表4展示了根据回归分析，用户的公平性与隐私取向及所处情境如何与收益相关属性的重要性感知相关联。

对于**公平性目标**，我们发现，捐赠数据意愿较强的用户（$\beta = 0.24, p < .001$）在排序任务中更倾向于优先考虑公平性目标，这表明他们密切关注自己的数据如何为社会公益作出贡献。隐私保护水平较高的用户则往往较少重视与数据捐赠相关的公平性目标。具体而言，用户对数据保护透明度重要性的感知与其在联合分析任务中对公平性目标的关注程度呈负相关（$\beta = -0.21, p \leq .05$）；而采取主动隐私保护措施的用户，在协议设计属性中往往将公平性目标排在较低位置（$\beta = -0.14, p \leq .05$）。

关于**金钱激励**，报告亲近的人遭受过歧视的个体，在不同协议设计之间进行选择时，不太可能优先考虑补偿福利（$\beta = -0.20, p < .01$）。稿件已提交至 ACM

非二元性别受访者在联合分析选择中对金钱激励的评价也较低（$\beta = -0.80, p \leq .05$）。此外，更重视数据使用透明度（$\beta = -0.20, p < .001$）并采取主动隐私保护措施（$\beta = -0.17, p < .001$）的个体，在对协议设计属性进行排序时，往往不太在意金钱激励。

表4. 用户的公平与隐私取向及情境与收益相关属性的相关性。“CBC”的回归基于从基于选择的联合分析中得出的个体层面属性重要性，该分析根据用户的选择揭示属性重要性。“排序”的回归基于根据用户直接属性排序得出的属性重要性，反映陈述属性重要性。模型间可靠性表示在CBC模型和排序模型中预测效应不相冲突的自变量所占百分比。***$p < 0.001$，**$p < 0.01$，*$p \leq 0.05$

用户的公平与隐私取向及情境	Coef（基于选择的联合分析）	标准误（基于选择的联合分析）	Coef（排序）	标准误差（排序）
公平性目标（模型间可靠性：85.19%）				
对数据保护透明度的感知重要性	-0.21*	0.10	-	-
总体数据捐赠意愿	-	-	0.24***	0.05
主动隐私保护措施	-	-	-0.15*	0.06
性别：女性（参照组：男性）	-0.07	0.07	-0.09*	0.05
金钱激励（模型间可靠性：85.19%）				
亲近关系中的感知歧视经历	-0.20**	0.07	-0.10	0.05
性别：非二元性别（参照组：男性）	-0.80*	0.36	-0.24	0.21
族裔：亚裔（参照组：白人）	0.24*	0.10	0.05	0.06
对数据使用透明度的感知重要性	-0.17	0.09	-0.20***	0.06
主动隐私保护措施	-0.10	0.09	-0.17***	0.05
教育程度：中等教育（参照：大学）	0.07	0.08	0.10*	0.05

6.2.2 注重隐私的用户更重视数据存储地点和隐私保护机制。通过回归分析，我们还确定了用户的公平与隐私取向及情境如何与对风险相关属性的感知重要性相关，如表5所示。

更重视数据使用透明度的参与者在其联合分析选择中对**隐私保护机制**赋予了更大权重（$\beta = 0.31,\ p < .01$）。同样，强调数据保护透明度的参与者也对这一机制给予了更高的重要性（$\beta = 0.24,\ p < .001$）。相比之下，在求职申请中经历过歧视的参与者相对于公平性监测协议的其他属性，对隐私保护机制赋予的重要性较低（$\beta = -0.29,\ p < .01$）。

最后，在**数据存储**方面，采取更主动隐私保护措施的参与者（$\beta = 0.24,\ p < .01$）以及对数据保护透明度感知重要性更强的参与者（$\beta = 0.24,\ p \leq .05$）更强调其数据的存储位置。与白人受访者相比，黑人受访者对数据存储的关注较少（$\beta = -0.33,\ p < .01$）。

6.2.3 对自动化招聘系统的理解会提高用户对收集的信息类型的关注，而边缘化群体则更重视数据使用方式。表6展示了用户的公平与隐私取向及其所处情境，如何与用户对涉及感知收益与风险的设计属性重要性的判断相关联，包括收集的信息类型、数据使用和监测主体。

对于**收集的信息类型**，对自动化招聘系统作用理解更深入的个体在联合分析决策中更关注这一属性（$\beta = 0.27, p < .001$）。与拥有大学教育背景的个体相比，接受基础教育的个体对联合分析任务中收集哪些类型的个人信息表现出明显更大的关注

表5. 用户的公平与隐私取向及情境与风险相关属性重要性的相关性。***$p < 0.001$, **$p < 0.01$, *$p \leq 0.05$

用户的公平与隐私取向及情境	Coef（基于选择的联合分析）	标准误（基于选择的联合分析）	Coef（排序）	标准误（排序）
隐私保护机制（模型间可靠性：85.19%）				
求职申请中的个人歧视经历	-0.29**	0.10	-0.08	0.05
对数据使用透明度的感知重要性	0.31**	0.10	-	-
对数据保护透明度的感知重要性	-	-	0.24***	0.05
性别：女性（参照组：男性）	0.11	0.07	0.10**	0.04
族裔：其他（参照：白人）	-0.09	0.12	-0.15*	0.06
数据存储（模型间可靠性：74.07%）				
族裔：黑人（参照：白人）	-0.33**	0.11	-0.04	0.06
主动隐私保护措施	0.24**	0.09	0.08	0.05
对数据保护透明度的感知重要性	0.24*	0.10	0.12	0.08

（$\beta = 0.69, p \leq .05$）。此外，既往数据捐赠经历正向预测了收集的信息类型在各属性中的较高排序（$\beta = 0.06, p < .01$）；而数据捐赠意愿较高的个体则认为该属性的重要性较低（$\beta = -0.15, p < .001$）。

对于**数据使用**，自我认同为黑人（$\beta = 0.24, p \leq .05$）、女性（$\beta = 0.17, p \leq .05$）或非二元性别（$\beta = 0.97, p < .01$）的受访者，更加重视其数据将如何被使用，且差异显著。这表明，历史上受边缘化群体会密切关注其数据是否会对算法公平性产生更直接的影响（评估算法，还是评估并开发算法）。相反，认为数据捐赠隐私风险较高的个体，不太可能将数据使用视为联合分析任务中的重要属性（$\beta = -0.19, p \leq .05$）。陈述的数据使用重要性与年龄较大（$\beta = -0.04, p \leq .05$）、受教育程度较低（$\beta = -0.12, p < .01$）以及在求职申请中遭受个人歧视的经历（$\beta = -0.10, p \leq .05$）呈负相关。

对于**监测主体**，非二元性别人士（$\beta = 0.74, p < .001$）和采取主动隐私保护措施的人（$\beta = 0.23, p < .001$）认为这一属性的重要性显著更高，反映出他们更加关注由谁来监测公平性。

主要发现

回归分析显示，愿意共享数据的用户往往优先考虑协议设计中的公平性目标。相比之下，注重隐私的用户表现出更积极的隐私行为，对数据保护透明度也更加敏感，因此较少重视公平性或金钱激励，而是关注数据存储和隐私保护机制等风险相关属性。此外，在选择偏好的协议时，对算法系统有更深入了解的用户会更加关注收集的信息类型。自我认同为黑人、女性或非二元性别的受访者，在联合分析任务中显著更加重视其数据将如何用于公平性监测。

7 讨论

本研究为用户如何评估并优先考虑公平性监测隐私保护协议的各项设计属性（RQ1），以及这些考量如何与用户的公平与隐私取向及情境相关（RQ2），提供了实证见解。本节将研究发现置于先前研究的背景下，并综合探讨其对未来隐私保护型公平性监测协议设计的启示。投稿至 ACM 的稿件

表6. 用户的公平与隐私取向及情境与收益-风险属性重要性的相关性。$^{***}p < 0.001$，$^{**}p < 0.01$，$^{*}p \leq 0.05$

用户的公平与隐私取向及情境	Coef（基于选择的联合分析）	标准误（基于选择的联合分析）	Coef（排序）	标准误（排序）
收集的信息类型（模型间可靠性：85.19%）				
对自动化招聘系统的理解	0.27**	0.08	-0.07	0.05
教育程度：基础教育（参照：大学）	0.69*	0.32	0.17	0.18
既往数据捐赠经历	-	-	0.06**	0.02
总体数据捐赠意愿	-0.07	0.07	-0.15***	0.04
数据使用（模型间可靠性：70.37%）				
族裔：黑人（参照：白人）	0.24*	0.11	-0.07	0.06
性别：女性（参照：男性）	0.17*	0.07	0.07	0.04
性别：非二元性别（参照：男性）	0.97**	0.36	-0.24	0.19
对数据捐赠的感知隐私风险	-0.19*	0.09	0.07	0.05
Age	-0.02	0.04	-0.04*	0.02
教育程度：中等教育（参照：大学）	-0.08	0.08	-0.12*	0.05
求职申请中的个人歧视	0.12	0.10	-0.10*	0.04
亲近关系中的感知歧视经历	-0.11	0.09	0.12*	0.05
监测主体（模型间可靠性：77.78%）				
族裔：其他（参照组：白人）	-0.38**	0.12	0.04	0.06
性别：非二元性别（参照：男性）	-	-	0.74**	0.21
主动隐私保护措施	-	-	0.23***	0.05

7.1 设计并传达保护隐私的公平性监测协议：以人为本的视角

7.1.1 公平性目标至关重要：使公平性目标与外展策略相契合，以构建具有代表性的数据集。正如第6.1节所示，公平性目标对用户在联合分析任务中的决策起着关键作用。值得注意的是，与男性相比，女性对群体公平的支持度更高，这与先前关于历史上占优势群体和处于弱势群体之间准确性与隐私权衡存在差异的研究发现一致 [75]。鉴于历史上处于弱势的群体在算法公平性研究所用数据集中持续代表性不足 [34], **将有针对性的外展与清晰传达公平性目标相结合**是构建更具代表性数据集的一种可行策略。不过，鉴于算法公平性对普通用户而言仍然较为抽象 [108], ，未来研究有必要探究有效的方法，以通俗易懂且引人入胜的方式传达公平性目标 [113]。

7.1.2 金钱激励的细微之处：激励参与，同时避免胁迫。研究发现，尽管在直接排序中，金钱激励被列为最不重要的属性，但在联合分析任务中，它对影响用户决策发挥了最重要的作用。因此，在假设情境中，金钱激励的影响远大于概念性情境。结合 Berke 的研究（该研究表明，与假设情境相比，金钱激励在现实情境中的作用甚至更强） [11], ，我们建议数据捐赠倡议中的利益相关方将透明的金钱激励视为重要的激励因素。不过，还需进行更细致的考量。研究发现表明，金钱激励对历史上占优势的群体（例如男性和较年轻的参与者）所产生的激励作用强于对历史上处于弱势的群体的作用，如第 6.1.2 节所示。先前研究 [6, 11] 也指出，对于具有内在激励的个体而言，金钱激励可能并非主要的参与驱动因素。因此，我们建议**将适度的金钱激励与对公平性目标的清晰说明相结合**，同时**避免过高的激励**

这可能带来胁迫风险，或因主要受经济收益驱动的参与者占比过高而扭曲样本构成。

7.1.3 并非越少越好，而是越清晰越好：关注数据与影响之间的联系，而非预先权衡。研究发现，要求提供较少数据或缩小数据使用范围（通常被认为能降低隐私风险），并不能可靠地提高接受度。具体而言，如第 6.1.2 节所示，与仅收集人口统计数据的协议相比，同时收集招聘决策和资历数据的协议获得了更广泛的接受。同样，参与者更倾向于将数据用于开发和评估，而非仅用于评估。这些研究发现表明，用户的隐私权衡取决于其感知到的影响和成效：当明确说明特定数据项如何带来具体的公平性改进，并辅以可信的保障措施和掌控感时，协议的信任度和接受度实际上可能会提升，而非降低。尽管数据最小化仍是一项重要原则 [42]，我们认为**当有必要扩大数据收集范围或拓展数据用途时，更有意义的做法是通过清晰说明数据与影响之间的联系，增进用户对其数据如何促成改变的理解，而不是预先在数据收集或使用方面作出取舍以促进参与**。为此，我们建议未来研究积极探索如何透明地说明数据生命周期 [81]，以阐明数据的价值和风险，支持知情同意。

7.1.4 迈向可信的 *MPC* 协议设计：采用可靠可信第三方的分布式存储。我们的研究发现，参与者对参与隐私保护型数据存储监测或协助工作的不同利益相关方抱有程度不一的信任。相较于由开发算法的公司开展内部监测，参与者更能接受由独立审计机构进行外部监测。同样，在协助分布式数据存储的第三方中，研究机构和非政府组织获得的信任显著高于专业商业实体。这些模式与先前关于用户对数据捐赠中中立或非利益相关方的信任研究相呼应，例如学术研究人员 [11]。在隐私保护机制方面，尽管匿名化和加密等基础方法已经获得较高的接受度，分布式存储进一步将接受度提高了约 10%，表明在多方计算协议中采用分布式存储进行公平性监测具有巨大潜力。鉴于多方计算协议的协调和实施固有的复杂性 [51]，，未来的数据捐赠倡议应**优先与用户信任的可信第三方合作**，例如研究机构或非政府组织，同时**确保具备支持安全且可扩展的 MPC 实施的技术基础设施**。

7.2 将协议接受度与用户背景及公平与隐私取向联系起来：定制多方计算协议沟通的启示

7.2.1 通过阐明贡献所体现的公平性相关价值，吸引历史上处于弱势的群体。我们发现，人口统计特征会影响个体如何在数据捐赠中权衡为算法公平性作出贡献与保护自身隐私。第 6.2.3 节显示，女性、黑人和非二元性别参与者对数据使用的关注显著更高。此外，第 6.1.2 节表明，与白人参与者相比，非白人参与者对不利于隐私的公平性监测设计接受度更高。这与第 6.2.2 节的研究发现一致：经历过更多求职歧视的受访者对隐私保护机制的关注较少。

既有研究表明，历史上处于弱势的群体对算法公平性抱有更高的预期 [75, 104, 108]。将这一研究领域拓展至数据捐赠情境后，我们的研究发现，来自历史上处于不利地位背景的个体往往优先考虑公平性目标，而非隐私顾虑，这为不同群体在数据捐赠考量方面的差异提供了

新的见解 [11, 88]。为此，我们呼吁设计能够适应不同人口群体预期的协议，以确保获得有效开展算法公平性测量所必需的包容性、具有代表性的数据集 [35]。此外，**透明地传达公平性相关价值，是促使历史上处于不利地位的群体参与其中的一种有前景的方法**，例如通过视觉对比展示提升公平性的效果。然而，这种方法应着重于客观告知，而非劝说用户凌驾于自身隐私顾虑之上。我们的目标是清晰阐明数据贡献如何推动实现公平性目标，从而帮助用户做出更知情的决策，同时尊重个人隐私偏好并确保真正的知情同意。

7.2.2 支持有意义同意的隐私权衡：根据用户取向调整沟通方式。 除了人口统计特征之外，我们的研究发现还揭示了公平与隐私取向如何影响隐私—收益权衡。第6.2.1节显示，隐私取向较强的参与者表现出更多隐私行为，也更关注数据使用和保护透明度，因此对社会或金钱收益的关注显著较少。相反，他们更关注隐私保护机制和所收集的信息类型等风险相关属性（第 6.2.2 节）。而数据分享意愿较强的参与者则更重视自己的贡献如何推动算法公平性。这些研究发现拓展了先前在接触者追踪应用采用研究中考察的隐私—收益权衡框架 [6, 44], ，该框架表明，初始接受度较高的用户更关注潜在收益，而持犹豫态度的用户则更强调风险。我们建议研究人员和从业者设计能够回应用户具体顾虑的自适应同意界面（例如，人工智能聊天机器人 [110]），**支持用户进行个性化探索，以便向其提供有意义的信息，介绍捐赠的益处或可用的隐私保护措施，同时避免技术细节过于复杂而令用户不堪重负**。值得注意的是，根据用户取向定制同意界面时，需要关注注意力焦点可能带来的偏差。为此，有必要向用户全面概述关键知识部分（例如隐私营养标签 [117]），并突出显示用户探索过程中可能忽略的信息，帮助他们了解此前未曾关注的方面。

7.3 反思陈述与显露的属性重要性之间的不一致

7.3.1 当信息的具体程度影响知情同意时：迈向透明、具体的协议沟通。本研究揭示了用户使用多方计算公平性监测协议时，陈述的属性重要性与显露的属性重要性之间存在不一致。例如，第6.1.1节显示，当用户在不了解属性如何发挥作用的情况下对属性进行排序时，他们往往会优先考虑隐私相关特征。然而，在评估完整的协议设计时，他们会将关注点转向金钱激励和公平性目标等益处。这一模式表明，当用户获得了关于隐私功能的更具体信息（例如确切的隐私保护机制和所收集的数据类型）时，他们的注意力更可能转向公平性贡献等益处。这些结果与 Farke 等人 [36], 的研究一致；该研究发现，感知到的透明度提高后，用户会更多地关注数据收集的积极方面。事实上，来自其他领域的大量证据表明，提供隐私相关信息并不会阻碍数据捐赠 [11, 39, 103]。尽管如此，如何降低透明度漂洗的风险，避免其助长不受约束的数据收集，仍是一个至关重要的议题 [36]。

本研究通过考察用户在不同风险认知水平和收益理解水平下如何以不同方式评估属性重要性，为隐私—收益权衡提供了新颖的视角。这些见解进一步凸显了在数据捐赠倡议中，为知情同意提供适当信息深度的重要性。**而不是**

仅传达“协议的关注重点”，并提供更细致、具体的协议设计细节，说明“协议实际上如何解决这些关注重点”，或许是促进知情同意的一种有效方式，使用户能够做出更准确反映其隐私—效益权衡中价值取向的选择。

7.3.2 结合陈述与显露的属性重要性，考察用户在不同设计方案中的优先事项。陈述与显露的属性重要性之间的不一致，与双加工理论相符 [33],。双加工理论将人类的思维与决策解释为两种不同认知系统共同作用的结果：系统1负责直觉加工，系统2负责审慎加工。一种可能的解释是，显露重要性（通过基于情境的比较，隐含地揭示属性对个人的重要性）更符合系统1，而陈述重要性（系统地比较抽象属性并对其重要性进行排序）则更能体现系统2。让用户参与不同认知过程，有助于更全面地了解用户如何评估技术协议设计，尤其是在复杂决策的情境下，例如本研究考察的隐私—收益权衡。因此，我们建议未来的研究人员和实践者**结合陈述与显露的属性重要性，考察用户在不同设计方案中的优先事项**，从而获得互补的见解。

7.4 局限性

本研究存在若干局限性。首先，我们在受《通用数据保护条例》约束的欧洲国家开展了调查；当地法律要求推动了符合数据保护要求的公平性监测协议的发展 [51, 101]。然而，公平性 [12] 和隐私 [68] 都是社会文化构念，在不同社会之间存在显著差异。因此，我们建议未来研究考察用户对隐私—效益权衡的看法，以指导不同文化背景下的隐私保护型公平性监测协议设计。其次，本研究采用定量方法，通过联合分析和回归模型考察用户对协议设计的接受度。访谈等定性研究可以进一步增进理解，揭示用户如何理解协议，以及如何形成符合其预期的偏好。第三，尽管联合分析任务能够逼真地模拟决策情境，但可能无法充分反映用户在实际数据捐赠情境中的认知与行为 [11]。为此，我们建议在现实场景部署期间持续收集用户反馈，以迭代改进并调整用于公平性监测的隐私保护协议。此外，尽管我们努力使参与者群体多元化，但非二元性别人士和年长求职者等一些边缘化群体在样本中仍然代表不足。由于这些群体在算法决策中可能面临不利处境，确保他们参与公平性监测至关重要。因此，开展更有针对性、更深入的研究，探讨这些群体对公平性监测数据共享的看法，将极具价值。最后，本研究关注算法招聘公平性监测情境下用户对多方计算协议的看法；部分研究发现可能无法推广至其他数据共享实践。我们建议未来研究考察用户的隐私—效益权衡应如何影响其他领域（如医疗、教育、移民等）的公平性监测协议设计。

8 结论

尽管数据捐赠为公平性监测提供了颇具前景的途径，但开发既符合用户对数据使用的预期、又能保障隐私保护的隐私保护协议，仍是一项重大挑战。为弥补这一不足，我们对欧洲的 833 名参与者开展了问卷调查研究，以考察用户对用于公平性监测的隐私保护协议设计的接受度。基于隐私权衡理论和实用的多方计算协议，稿件已提交至 ACM

针对公平性监测，我们设计了与收益相关、与风险相关，以及同时涉及感知收益与风险的设计属性，以供研究。

首先，我们通过属性排序得出的陈述重要性，以及基于选择的联合分析任务得出的揭示重要性，评估了设计属性重要性。我们发现，用户在直接评估中优先考虑隐私相关特征（所收集的信息类型和隐私保护机制），但在模拟决策情境中做出选择时，则更重视收益相关属性（公平性目标和金钱激励）。用户的隐私价值观体现在其偏好的协议设计中，包括相较于基础匿名化和加密，更强烈地偏好分布式数据存储。然而，用户却矛盾地表现出对请求更多数据和更广泛数据使用的协议更高的接受度，这可能是因为他们更强烈地感到自己在为公平性做出贡献。这些收益-隐私权衡在不同人口统计群体之间存在差异。例如，非白人参与者在隐私条件不太有利的情况下，更愿意为公平性监测共享数据。此外，回归分析揭示了协议接受度与用户的公平与隐私取向及情境之间的相关性。具有初始数据捐赠意愿的用户更强调公平性目标，而注重隐私的用户，即那些具有积极隐私行为且对数据保护透明度高度敏感的用户，则不太重视公平性或金钱激励，而是更关注隐私功能。

基于这些研究发现，我们提出了隐私保护型公平性监测协议的实践启示，涵盖符合用户期望的包容性协议设计，以及能够实现有意义知情同意的有效协议沟通。

致谢

本工作由 FINDHR 项目、地平线欧洲计划资助，资助协议编号：101070212。

参考文献

[1] Nitin Agrawal、Reuben Binns、Max Van Kleek、Kim Laine 和 Nigel Shadbolt。《隐私保护计算开发中的设计与治理挑战探索》。载于*2021*年 *CHI* 人机交互大会论文集，第 1–13 页，2021 年。[2] Tawfiq Alashoor、Mark Keil、Zhenhui Jack Jiang 和 Kambiz Saffarizadeh。《隐私顾虑与数据捐赠：社会效益是否重要？》《*MIS*季刊》，49(2)，2025 年。[3] Sandro Ambuehl、Muriel Niederle 和 Alvin E Roth。《更多的钱，更多的问题？高薪是否可能具有强迫性且令人反感？》《美国经济评论》，105(5):357–360，2015 年。[4] Arno Appenzeller、NickTerzer、Erik Krempel 和 Jürgen Beyerer。《利用隐私保护技术实现私人医疗数据捐赠》。载于第 *15* 届普适技术与辅助环境国际会议论文集，第 446–454 页，2022 年。[5] Carolyn Ashurst 和 Adrian Weller。《没有人口统计数据的公平性：方法综述》。载于第 *3* 届 *ACM*算法、机制与优化公平及可及性会议论文集，第 1–12 页，2023 年。[6] Oshrat Ayalon、Dana Turjeman 和 Elissa M Redmiles。《探索新冠肺炎{接触者追踪应用} 采纳中的隐私与激励考量》。载于第 *32* 届 *USENIX*安全研讨会（*USENIX Security 23*），第 517–534 页，2023 年。[7] Susanne E Baumgartner、Sindy R Sumter、Vladislav Petkevič 和 Wisnu Wiradhany。《一种新颖的 iOS 数据捐赠方法：纵向研究中的自动处理、依从性与反应性》。《社会科学计算评论》，41(4):1456–1472，2023 年。[8] Rachel KE Bellamy、Kuntal Dey、Michael Hind、Samuel C Hoffman、Stephanie Houde、Kalapriya Kannan、Pranay Lohia、Jacquelyn Martino、Sameep Mehta、Aleksandra Mojsilović 等。《AI Fairness 360：用于检测和缓解算法偏见的可扩展工具包》。《*IBM*研究与开发杂志》，63(4/5):4–1，2019 年。[9] Florian Bemmann 和 Sven Mayer。《数据隐私对用户智能手机应用采纳决策的影响》。《*ACM*人机交互论文集》，8(MHCI):1–23，2024 年。[10] Yoav Benjamini 和 Yosef Hochberg。《控制假发现率：一种实用且有效的多重检验方法》。《英国皇家统计学会杂志：*B*辑（方法论）》，57(1):289–300，1995 年。[11] Alex Berke、Robert Mahari、Alex Pentland、Kent Larson 和 Dana Calacci。《从数千名美国亚马逊用户众包数据实验中获得的见解：透明度、金钱与数据使用的重要性》。《*ACM*人机交互论文集》，8(CSCW2):1–48，2024 年。

[12] John J Berman、Virginia Murphy-Berman 和 Purnima Singh。《对公平感知的跨文化相似性与差异》。《跨文化心理学杂志》，16(1):55–67，1985 年。[13] Surabhi Bhargava 和 Shubham Singhal。《机器学习模型部署后监测中的挑战、解决方案与最佳实践》。[14] Matthew Bietz、Kevin Patrick 和 Cinnamon Bloss。《数据捐赠：一种公民科学健康研究模式》。《公民科学：理论与实践》，4(1)，2019 年。[15] Laura Boeschoten、Jef Ausloos、Judith E Möller、Theo Araujo 和 Daniel L Oberski。《通过数据捐赠收集隐私保护数字踪迹数据的框架》。《计算传播研究》，4(2):388–423，2022 年。[16] Anne Bowser、Katie Shilton、Jenny Preece 和 Elizabeth Warrick。《将隐私纳入公民科学考量：开放环境中的伦理研究》。载于《2017 年 ACM计算支持的协同工作与社会计算会议论文集》，第 2124–2136 页，2017 年。[17] Stephen Breen、Karim Ouazzane 和 Preeti Patel。《《通用数据保护条例》：你的同意有效吗？》。《商业信息评论》，37(1):19–24，2020 年。[18] Johannes Breuer、Zoltán Kmetty、Mario Haim 和 Sebastian Stier。《后API时代收集 Facebook 数据的以用户为中心的方法：两项研究的经验与未来研究建议》。《信息、传播与社会》，26(14):2649–2668，2023 年。[19] 波士顿女性劳动力委员会（BWWC）。《数据隐私》。https://thebwwc.org/mpc。访问日期：2025-11-08。[20] Toon Calders 和 Sicco Verwer。《三种实现无歧视分类的朴素贝叶斯方法》。《数据挖掘与知识发现》，21(2):277–292，2010 年。[21] Simon Caton 和 Christian Haas。《机器学习公平性：综述》。《ACM计算综述》，56(7):1–38，2024 年。[22] Hongyan Chang 和 Reza Shokri。《算法公平性的隐私风险》。载于《2021 年 IEEE欧洲安全与隐私研讨会（EuroS&P）》，第 292–303 页。IEEE，2021 年。[23] Victoria Cheng、Vinith M Suriyakumar、Natalie Dullerud、Shalmali Joshi 和 Marzyeh Ghassemi。《弄假成真？差分隐私合成数据对下游分类公平性的影响》。载于《2021 年 ACM公平性、问责制与透明度会议论文集》，第 149–160 页，2021 年。[24] Mary J Culnan 和 Pamela K Armstrong。《信息隐私担忧、程序公平与非人际信任：一项实证调查》。《组织科学》，10(1):104–115，1999 年。[25] Rachel Cummings、Gabriel Kaptchuk 和 Elissa M Redmiles。《“我需要更好的描述”：对用户差分隐私期望的调查》。载于《2021 年 ACM SIGSAC计算机与通信安全会议论文集》，第 3037–3052 页，2021 年。[26] Marco De Vincenzi、Ilaria Matteucci、Fabio Martinelli 和 Stefano Sebastio。《安全两方计算在隐私保护 Android 应用中的应用》。载于《第 18 届国际可用性、可靠性与安全会议论文集》，第 1–7 页，2023 年。[27] Oscar Blessed Deho、Lin Liu、Jiuyong Li、Jixue Liu、Chen Zhan 和 Srecko Joksimovic。《当过去 = 未来：评估数据集漂移对学习分析模型公平性的影响》。《IEEE学习技术汇刊》，17:1007–1020，2024 年。[28] Onyinye Dibia、Brad Stenger、Steven Baldasty、Mako Bates、Ivoline C Ngong、Yuanyuan Feng 和 Joseph P Near。《差分隐私中的可用性研究：系统化综述》。《arXiv》预印本 arXiv:2412.16825，2024 年。[29] Lilian Edwards。《欧盟《人工智能法案》：其重要性与适用范围概述》。《人工智能（欧盟《人工智能法案》）》，1:25，2021 年。[30] Felix Eggers 和 Henrik Sattler。《混合个体化两层选择型联合分析（HIT-CBC）：一种测量具有众多属性水平的偏好结构的新方法》。《国际营销研究杂志》，26(2):108–118，2009 年。[31] Zachary Espiritu、Marilyn George、Seny Kamara 和 Lucy Qin。《SynQ：针对加密数据的公共政策分析》。载于《2024 年 IEEE安全与隐私研讨会（SP）》，第 146–165 页。IEEE，2024 年。[32] 欧盟委员会。《2018年欧盟数据保护规则改革》。可访问：https://ec.europa.eu/commission/sites/beta-political/files/data- protection-factsheet-changes_en.pdf，2018 年。[33] Jonathan St BT Evans 和 Keith E Stanovich。《高级认知的双加工理论：推进这场辩论》。《心理科学展望》，8(3):223–241，2013 年。[34] Alessandro Fabris、Nina Baranowska、Matthew J Dennis、David Graus、Philipp Hacker、Jorge Saldivar、Frederik Zuiderveen Borgesius 和 Asia J Biega。《算法招聘中的公平性与偏见：多学科综述》。《ACM智能系统与技术汇刊》，2024 年。[35] Alessandro Fabris、Stefano Messina、Gianmaria Silvello 和 Gian Antonio Susto。《算法公平性数据集：迄今为止的发展历程》。《数据挖掘与知识发现》，36(6):2074–2152，2022 年。[36] Florian M Farke、David G Balash、Maximilian Golla、Markus Dürmuth 和 Adam J Aviv。《隐私仪表板对最终用户有益吗？评估用户对 Google“我的活动”的感知与反应》。载于《第 30 届 USENIX安全研讨会（USENIX Security 21）》，第 483–500 页，2021 年。[37] Victoria Fast 和 Daniel Schnurr。《用于数字接触者追踪的数据捐赠：金钱激励的短期与长期影响》。可见 SSRN 3786245，2023 年。[38] Emilio Ferrara。《人工智能中的公平性与偏见：来源、影响及缓解策略简述》。《Sci》，6(1):3，2023 年。[39] Daniel Franzen、Claudia Müller-Birn 和 Odette Wegwarth。《传达隐私—效用权衡：利用差分隐私隐私决策界面支持知情的数据捐赠》。《ACM人机交互论文集》，8(CSCW1):1–56，2024 年。[40] Markus Freitag 和 Julian Schuessler。《cjpowr——联合分析实验的先验功效分析》。R软件包，2020 年。[41] Sorelle A Friedler、Carlos Scheidegger 和 Suresh Venkatasubramanian。《公平性是否可能？不同的价值体系需要不同的公平决策机制》。《ACM通讯》，64(4):136–143，2021 年。[42] Prakhar Ganesh、Cuong Tran、Reza Shokri 和 Ferdinando Fioretto。《机器学习中的数据最小化原则》。载于《2025 年 ACM公平性、问责制与透明度会议论文集》，第 3075–3093 页，2025 年。稿件已提交至 ACM

[43] Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault 和 Thomas W Malone。人类与大型语言模型交互模式分类：初步探索。载于《CHI人机交互大会扩展摘要》，第 1–11 页，2024 年。[44] Sarah Geber 和 Thomas N Friemel。基于类型学的方法追踪新冠肺炎疫情期间追踪应用的采用情况：以 SwissCovid 应用为例。《定量描述期刊》，1：1–28，2021 年。[45] Reza Ghaiumy Anaraky、Yao Li、Hichang Cho、Danny Yuxing Huang、Kaileigh Angela Byrne、Bart Knijnenburg 和 Oded Nov。依据个体的调节焦点实现个性化隐私保护：你会维护还是增强自己的信息隐私？载于《2024年CHI人机交互大会论文集》，第 1–17 页，2024 年。[46] Oded Goldreich。安全多方计算。稿件。初步版本，78(110)：1–108，1998 年。[47] Alejandra Gomez Ortega、Jacky Bourgeois、Wiebke Toussaint Hutiri 和 Gerd Kortuem。超越数据交易：实现知情且有意义的数据捐赠框架。《人工智能与社会》，第 1–18 页，2023 年。[48] Paul E Green 和 Vithala R Rao。联合测量——量化判断数据。《营销研究杂志》，8(3)：355–363，1971 年。[49] Lara Groves、Jacob Metcalf、Alayna Kennedy、Briana Vecchione 和 Andrew Strait。审计工作：探索纽约市的算法偏见审计制度。载于《2024年ACM公平性、问责制与透明度会议论文集》，第 1107–1120 页，2024 年。[50] Changyang He、Yue Deng、Alessandro Fabris、Bo Li 和 Asia Biega。基于求职者公平性顾虑构建公平的在线招聘框架。arXiv 预印本 arXiv:2501.14110，2025 年。[51] Lukas Helminger 和 Christian Rechberger。GDPR 中的多方计算。载于《隐私研讨会：数据保护法的国际趋同与创新技术合规》，第 21–39 页。Springer，2022 年。[52] Kirsten Hillebrand、Lars Hornuf、Benjamin Müller 和 Daniel Vrankar。大数据的社会困境：捐赠个人数据以促进社会福利。《信息与组织》，33(1)：100452，2023 年。[53] Kenneth Holstein、Jennifer Wortman Vaughan、Hal Daumé III、Miro Dudik 和 Hanna Wallach。提升机器学习系统的公平性：行业从业者需要什么？载于《2019年CHI人机交互大会论文集》，第 1–16 页，2019 年。[54] Chris Jay Hoofnagle、Bart Van Der Sloot 和 Frederik Zuiderveen Borgesius。欧盟《通用数据保护条例》：其内容及意义。《信息与通信技术法》，28(1)：65–98，2019 年。[55] Max Hort、Zhenpeng Chen、Jie M Zhang、Mark Harman 和 Federica Sarro。机器学习分类器的偏见缓解：全面综述。《ACM负责任计算期刊》，1(2)：1–52，2024 年。[56] Anna Lena Hunkenschroer 和 Alexander Kriebitz。人工智能招聘是否合乎伦理？从人权视角看人工智能在招聘中的应用。《人工智能与伦理》，3(1)：199–213，2023 年。[57] Rashidul Islam、Kamrun Naher Keya、Shimei Pan、Anand D Sarwate 和 James R Foulds。差异公平性：一种实现公平人工智能的交叉性框架。《熵》，25(4)：660，2023 年。[58] Matthew Jagielski、Michael Kearns、Jieming Mao、Alina Oprea、Aaron Roth、Saeed Sharifi-Malvajerdi 和 Jonathan Ullman。差分隐私公平学习。载于《国际机器学习会议》，第 3000–3008 页。PMLR，2019 年。[59] Seiie Jang、Eun-Jung Ko 和 Woontack Woo。统一的以用户为中心的情境：谁、何地、何时、何事、如何以及为何。载于 ubiPCMM，2005 年。[60] Bailey Kacsmar、Vasisht Duddu、Kyle Tilbury、Blase Ur 和 Florian Kerschbaum。从混乱中理解：迈向隐私计算的知情同意。载于《2023年ACM SIGSAC计算机与通信安全会议论文集》，第 210–224 页，2023 年。[61] Zoltán Kmetty、Ádám Stefkovics、Júlia Számely、Dongning Deng、Anikó Kellner、Edit Pauló、Elisa Omodei 和 Júlia Koltai。决定个体是否愿意捐赠社交媒体平台数据的因素。《信息、传播与社会》，第 1–26 页，2024 年。[62] Nima Kordzadeh 和 Maryam Ghasemaghaei。算法偏见：综述、综合与未来研究方向。《欧洲信息系统期刊》，31(3)：388–409，2022 年。[63] Patrick Kühtreiber、Viktoriya Pak 和 Delphine Reinhardt。复现研究：差分隐私相关说明对德国用户理解程度和数据共享态度的影响。载于《第十八届可用隐私与安全研讨会（SOUPS 2022）》，第 117–134 页，2022 年。[64] Ali Ladak、Jamie Harris 和 Jacy Reese Anthis。人们最关心哪些人工智能？一项关于道德考量的联合分析实验。载于《CHI人机交互大会论文集》，第 1–11 页，2024 年。[65] Andrei Lapets、Frederick Jansen、Kinan Dak Albab、Rawane Issa、Lucy Qin、Mayank Varia 和 Azer Bestavros。便捷的隐私保护型网页数据分析：评估并解决经济不平等问题。载于《第1届ACM计算与可持续社会SIGCAS会议论文集》，第 1–5 页，2018 年。[66] Tai Le Quy、Arjun Roy、Vasileios Iosifidis、Wenbin Zhang 和 Eirini Ntoutsi。公平性感知机器学习数据集综述。《Wiley跨学科评论：数据挖掘与知识发现》，12(3)：e1452，2022 年。[67] He Li、Jing Wu、Yiwen Gao 和 Yao Shi。考察个体对医疗可穿戴设备的采用：隐私权衡视角下的实证研究。《国际医学信息学杂志》，88：8–17，2016 年。[68] Yao Li、Eugenia Ha Rim Rho 和 Alfred Kobsa。情境因素和隐私顾虑对用户在社交网站上隐私决策的影响：文化差异。《行为与信息技术》，41(3)：655–677，2022 年。[69] Hangcheng Liu。方差不齐情况下 Welch 方差分析、Kruskal-Wallis 检验与传统方差分析的比较。弗吉尼亚联邦大学，2015 年。[70] Qinyi Liu、Oscar Deho、Farhad Vadiee、Mohammad Khalil、Srecko Joksimovic 和 George Siemens。合成数据能否兼顾公平性与隐私？合成数据生成方法与公平性算法的比较研究。arXiv 预印本 arXiv:2501.01785，2025 年。稿件已提交 ACM

[71] Yong Liu、Denzil Ferreira、Jorge Goncalves、Simo Hosio、Pratyush Pandab 和 Vassilis Kostakos。向科学研究捐赠情境数据：社会信号和感知对行动的影响。《与计算机交互》，29(2)：132–146，2017 年。[72] Jie Lu、Anjin Liu、Fan Dong、Feng Gu、Joao Gama 和 Guangquan Zhang。概念漂移下的学习：综述。《IEEE知识与数据工程汇刊》，31(12)：2346–2363，2018 年。[73] 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。[74] Philip Menard 和 Gregory J Bott。人工智能滥用与信息隐私顾虑：新的构念验证及未来方向。《信息系统期刊》，2024 年。[75] Mehdi Mourali、Dallas Novakowski、Ruth Pogacar 和 Neil Brigden。美国公众对算法决策中准确性与公平性权衡的看法。PLOS ONE，20(3)：e0319861，2025 年。[76] Priyanka Nanayakkara、Mary Anne Smart、Rachel Cummings、Gabriel Kaptchuk 和 Elissa M Redmiles。概率有多大？解释差分隐私中的 epsilon 参数。载于《第32届 USENIX 安全研讨会（USENIX Security 23）》，第 1613–1630 页，2023 年。[77] Henni Parviainen。欧盟的算法招聘系统能否依法使用自动化决策?《欧洲劳动法期刊》，13(2)：225–248，2022 年。[78] Mayana Pereira、Meghana Kshirsagar、Sumit Mukherjee、Rahul Dodhia、Juan Lavista Ferres 和 Rafael de Sousa。评估端到端表格数据机器学习流水线中的差分隐私合成数据的效用与公平性。PLOS ONE，19(2)：e0297271，2024 年。[79] Nico Pfiffner 和 Thomas N Friemel。利用数据捐赠开展传播学研究：探究影响捐赠意愿的因素。《传播方法与测量》，17(3)：227–249，2023 年。[80] Benny Pinkas、Thomas Schneider、Nigel P Smart 和 Stephen C Williams。安全两方计算具有实用性。载于《密码学进展——ASIACRYPT 2009：第15届密码学理论与应用国际会议，日本东京，2009年12月6日至10日，会议论文集15》，第 250–267 页。Springer，2009 年。[81] Neoklis Polyzotis、Sudip Roy、Steven Euijong Whang 和 Martin Zinkevich。生产环境机器学习中的数据生命周期挑战：综述。ACM Sigmod Record，47(2)：17–28，2018 年。[82] Yu Pu 和 Jens Grossklags。评估{朋友们的}隐私：共享个人数据时，匿名性是否重要？载于《第十三届可用隐私与安全研讨会（SOUPS 2017）》，第 339–355 页，2017 年。[83] Lucy Qin、Andrei Lapets、Frederick Jansen、Peter Flockhart、Kinan Dak Albab、Ira Globus-Harris、Shannon Roberts 和 Mayank Varia。从可用性到安全计算，再回归可用性。载于《第十五届可用隐私与安全研讨会（SOUPS 2019）》，第 191–210 页，2019 年。[84] Manish Raghavan、Solon Barocas、Jon Kleinberg 和 Karen Levy。缓解算法招聘中的偏见：评估相关主张与实践。载于《2020年公平性、问责制与透明度会议论文集》，第 469–481 页，2020 年。[85] Bashir Rastegarpanah、Mark Crovella 和 Krishna P Gummadi。公平输入与公平输出：隐私公平性与准确性公平性之间的不相容性。载于《第28届ACM用户建模、自适应与个性化会议附属出版物》，第 260–267 页，2020 年。[86] Afsaneh Razi、Ashwaq AlSoubai、Seunghyun Kim、Nurun Naher、Shiza Ali、Gianluca Stringhini、Munmun De Choudhury 和 Pamela J Wisniewski。Instagram 数据捐赠：一项案例研究，探讨如何收集生态效度良好的社交媒体数据，用于青少年在线风险检测。载于《CHI人机交互大会扩展摘要》，第 1–9 页，2022 年。[87] Gesine Richter、Christoph Borzikowsky、Bimba Franziska Hoyer、Matthias Laudes 和 Michael Krawczak。个人医疗数据的二次研究使用：患者对数据捐赠的态度。《BMC 医学伦理学》，22：1–10，2021 年。[88] Bence Ságvári、Attila Gulyás 和 Júlia Koltai。对参与被动数据收集实验的态度。《Sensors》，21(18)：6085，2021 年。[89] Julian Schuessler 和 Markus Freitag。联合分析实验的功效分析。2020 年。[90] John S Seberger 和 Sameer Patil。我们、他们（以及它）：社会取向、隐私顾虑以及美国公众对疫情追踪应用的预期使用情况。载于《2021年CHI人机交互大会论文集》，第 1–19 页，2021 年。[91] Avital Shulner-Tal、Tsvi Kuflik 和 Doron Kliger。公平性、可解释性及其间：理解不同解释方法对非专业用户关于算法系统公平性感知的影响。《伦理与信息技术》，24(1)：2，2022 年。[92] Henning Silber、Johannes Breuer、Christoph Beuthner、Tobias Gummer、Florian Keusch、Pascal Siegers、Sebastian Stier 和 Bernd Weiß。关联调查与数字踪迹数据：两项关于数据共享行为决定因素的研究发现。《英国皇家统计学会会刊A辑：统计学会刊》，185（增刊_2）：S387–S407，2022 年。[93] Anya Skatova 和 James Goulding。个人数据捐赠心理学。PLOS ONE，14(11)：e0224240，2019 年。[94] Anya Skatova、Esther Ng 和 James Goulding。数据捐赠：为公共利益共享个人数据。Application of Digital Innovation。英格兰伦敦：N-Lab，2014 年。[95] Joanna Sleigh。向心理健康研究捐赠个人数据的经历：一项探索性人类学研究。《Biomedical Informatics Insights》，10：1178222618785131，2018 年。[96] Theresa Stadler 和 Carmela Troncoso。隐私保护型数据共享机制为何难以实现。《Nature Computational Science》，2(4)：208–210，2022 年。[97] 欧盟。第10条：数据与数据治理。https://artificialintelligenceact.eu/article/10/。访问日期：2025-08-25。[98] 欧盟。第72条：提供者的上市后监测以及高风险人工智能系统的上市后监测计划。https: //artificialintelligenceact.eu/article/72/。访问日期：2025-08-25。[99] Marvin van Bekkum。利用敏感数据消除人工智能系统偏见：欧盟《人工智能法案》第10(5)条。arXiv 预印本 arXiv:2410.14501，2024 年。稿件已提交 ACM

[100] Michael Veale 和 Reuben Binns。现实世界中更公平的机器学习：在不收集敏感数据的情况下减轻歧视。*Big Data &Society*，4(2):2053951717743530，2017。[101] Paul Voigt 和 Axel Von dem Bussche。欧盟《通用数据保护条例》（GDPR）。实用指南，第*1*版，*Cham*：*Springer International Publishing*，10(3152676):10–5555，2017。[102] Emma Walquist、Isha Datey、Wenqi Zheng、Xiangyu Zhou、Kelly Berishaj、Melissa McDonald、Michele Parkhill、Dongxiao Zhu 和 Douglas Zytko。集体同意：捐赠代表多人的数据需要谁的同意？2025。[103] Ruizhe Wang、Roberta De Viti、Aarushi Dubey 和 Elissa M Redmiles。隐私保障在为利他目标自愿捐赠私人数据中的作用。*arXiv* 预印本 *arXiv:2407.03451*，2024。[104] Ruotong Wang、F Maxwell Harper 和 Haiyi Zhu。影响算法决策中感知公平性的因素：算法结果、开发流程与个体差异。载于*2020*年 *CHI* 人机交互大会论文集，第1–14页，2020。[105] Tien Wang、Trong Danh Duong 和 Charlie C Chen。通过移动应用披露个人信息的意向：隐私权衡视角。*International Journal of Information Management*，36(4):531–542，2016。[106] Alexander Wenz、Annette Jäckle 和 Mick P Couper。在概率抽样住户面板中使用移动技术进行数据收集的意愿。载于*SurveyResearchMethods*，第13卷，第1–22页，2019。[107] Cedric Deslandes Whitney 和 Justin Norman。虚假数据的真实风险：合成数据、多样性漂洗与规避同意。载于*2024*年 *ACM*公平性、问责制与透明度会议论文集，第1733–1744页，2024。[108] Allison Woodruff、Sarah E Fox、Steven Rousso-Schindler 和 Jeffrey Warshaw。对算法公平性认知的定性探索。载于*2018*年 *CHI* 人机交互大会论文集，第1–14页，2018。[109] Allison Woodruff、Vasyl Pihur、Sunny Consolvo、Laura Brandimarte 和 Alessandro Acquisti。隐私原教旨主义者会以1000美元出售自己的DNA吗……如果结果不会有任何坏处？韦斯廷分类、行为意向与后果。载于第*10*届可用隐私与安全研讨会（*SOUPS 2014*），第1–18页，2014。[110] Ziang Xiao、Tiffany Wenting Li、Karrie Karahalios 和 Hari Sundaram。让不知情者知情：利用人工智能聊天机器人提升用户阅读在线知情同意内容的体验。载于*2023*年 *CHI* 人机交互大会论文集，第1–17页，2023。[111] Aiping Xiong、Tianhao Wang、Ninghui Li 和 Somesh Jha。实现有效的差分隐私沟通，助力用户做出数据共享决策并加深理解。载于*2020*年 *IEEE*安全与隐私研讨会（*SP*），第392–410页。IEEE，2020。[112] Aiping Xiong、Chuhao Wu、Tianhao Wang、Robert W Proctor、Jeremiah Blocki、Ninghui Li 和 Somesh Jha。利用插图传达差分隐私信任模型：对用户理解、感知和数据共享决策的调查。*arXiv* 预印本 *arXiv:2202.10014*，2022。[113] Xinyuan Yan、Youjia Zhou、Arul Mishra、Himanshu Mishra 和 Bei Wang。探索人工智能教育中的公平性可视化。载于*2024*年 *IEEE* 第*17*届太平洋可视化会议（*PacificVis*），第1–10页。IEEE，2024。[114] Juliette Zaccour、Reuben Binns 和 Luc Rocher。访问被拒：为定量算法审计提供有意义的数据访问。载于*2025*年 *CHI* 人机交互大会论文集，第1–31页，2025。[115] Savvas Zannettou、Olivia Nemes-Nemeth、Oshrat Ayalon、Angelica Goetzen、Krishna P Gummadi、Elissa M Redmiles 和 Franziska Roesner。利用数据捐赠分析用户对 TikTok 短视频推荐的参与度。载于*2024*年 *CHI* 人机交互大会论文集，第1–16页，2024。[116] Wenqi Zheng、Emma Walquist、Isha Datey、Xiangyu Zhou、Kelly Berishaj、Melissa Mcdonald、Michele Parkhill、Dongxiao Zhu 和 Douglas Zytko。“这不是我们原本想要探讨的内容，但我觉得或许应该关注”：借助在线交友性暴力数据捐赠平台，学习如何进行创伤知情设计。载于*2024*年 *CHI* 人机交互大会论文集，第1–15页，2024。[117] Verena Zimmermann、Adrienn Toth、Hannah Sievers、Linda Fanconi、Yanis Isenring、Mona Henz、Alina Stöver 和 Nina Gerber。让我们以视觉呈现：测试用于传达隐私政策和数据处理信息的视觉类比与隐喻。载于*2025*年 *IEEE*安全与隐私研讨会（*SP*），第2303–2321页。IEEE，2025。[118] Tomasz Zukowski 和 Irwin Brown。考察人口统计因素对互联网用户信息隐私顾虑的影响。载于*2007*年南非计算机科学家与信息技术人员学会发展中国家 *IT* 研究年度研究会议论文集，第197–204页，2007。

效用估计的详细解释

参与者 i 在任务 t 中选择选项 j 的概率定义为：

$$P_{ijt} = \frac{e^{V_{ijt}}}{\sum_{k \in J} e^{V_{ikt}}}$$

其中，总效用 U_{ijt} 由系统部分 V_{ijt} 和误差项组成：

$$U_{ijt} = V_{ijt} + \epsilon_{ijt}, \quad \text{with} \quad V_{ijt} = \beta_0 + \sum_a \beta_a x_{ajt}.$$

此处， β_a 表示属性 a的效用权重， x_{ajt} 表示任务 t中选项 j的属性 a 水平， ϵ_{ijt} 则反映未观测方差。我们为每位受访者计算了以零为中心的效用值（在各属性水平上的总和为零）和属性重要性得分，每位受访者的得分总和为 100%。各属性的相对重要性计算如下：

$$I_a = \frac{\max(U_a) - \min(U_a)}{\sum_b (\max(U_b) - \min(U_b))}$$

这种方法使我们能够识别不同的设计属性如何影响用户对协议设计的决策。

B 调查详情

B.1 背景信息

- **什么是数据捐赠?**
  - 数据捐赠是指**自愿分享个人数据**以用于特定目的 .
  - 它可以支持**自动化招聘系统中的公平性**。

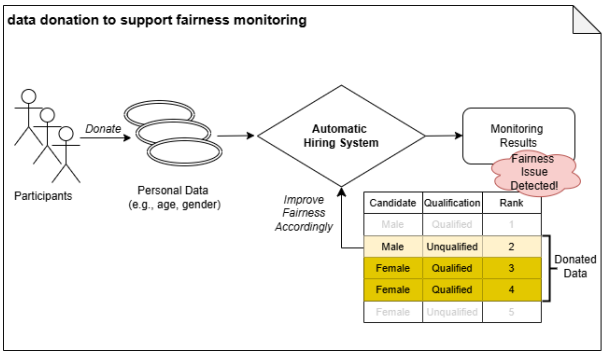


图 4. 数据捐赠以支持公平性监测

- **什么是自动化招聘系统?** 通过以下方式帮助招聘人员：– 自动审核求职申请– 更有效地推荐最佳候选人

- **为什么公平性在自动化招聘系统中很重要?** – 自动化招聘系统可能**继承偏差**，这些偏差来自其训练数据。– 公平性确保招聘决策**无偏见且不具歧视性**。– 捐赠数据可以帮助**减少偏差并提高公平性**。

- **什么是个人数据?** – 任何可用于**将您识别为个人的信息** – 有关您的**身体或文化身份的详细信息**

- **什么是敏感数据?**

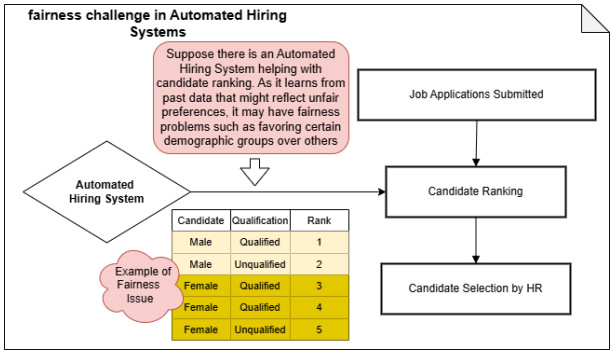


图 5. 自动化招聘系统中的公平性挑战

– 敏感数据是个人数据的一个子集，受到**更严格的法律保护**。

- **为什么保护自动化招聘系统中的个人数据很重要？** – 可防止 **候选人数据**遭到滥用。– 通过 **在第三方和服务提供商之间拆分数据**，任何一方都无法单独访问敏感信息。

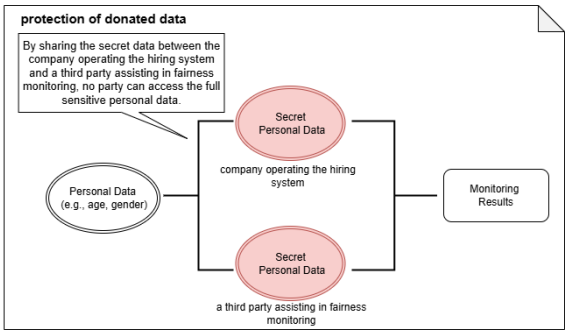


图 6. 捐赠数据的保护

B.2 评估问题

阅读完概述后，请回答以下问题。[3]

- 自动化招聘系统总是能做出公平的决定。（是/否）
- 数据捐赠有助于监测公平性问题（是/否）
- 存在一种公平性监测解决方案，第三方可以在无人存储求职候选人的完整个人数据的情况下监测公平性。（是/否）

B.3 人口统计问题

- 您的年龄是多少？ [未满18岁/18-24岁/25-34岁/35-44岁/45-54岁/55-64岁/65+]

[3]正确答案：否，是，是

- 您的性别是什么？ [男性/女性/非二元性别/不愿透露/其他]
- 您的族裔是什么（可多选） [白人（例如，欧洲、中东、北非）/黑人（例如，非洲、非裔加勒比）/亚裔（例如，南亚、东亚、东南亚）/混合或多个族群/不愿透露/其他]

- 您的最高教育水平是什么？ [未接受正规教育/初等教育/初中教育（例如，GCSE、初级证书）/高中教育（例如，A-levels、中学毕业会考、离校证书）/职业或技术培训/学士学位（或同等学历）/硕士学位（或同等学历）/博士学位（PhD 或同等学历）/其他]

- 您的就业状况如何？[就业/失业/不愿透露]
- 您曾为研究或社会公益捐赠数据多少次？ [从未/1 次/2–3 次/4–5 次/超过 5 次]

B.4 李克特量表问题

- 您个人在任何情境下经历歧视的程度如何？ [-3（完全没有）- 3（程度非常高）]

- 您认为自己在求职过程中经历歧视的程度如何？ [-3（完全没有）至 3（程度非常高）]

- 您身边亲近的人（例如家人、朋友、亲戚）在求职时经历歧视的程度如何？ [-3（完全没有）至 3（程度非常高）]
- 您对在线分享个人信息可能带来的风险和后果了解多少？ [-3（完全不了解）至 3（非常了解）]

- 您对日常使用的数字服务如何收集、使用、共享和保护您的个人数据了解多少？ [-3（完全不了解）至 3（非常了解）]
- 您多常会主动采取措施保护自己在网上的个人信息？ [-3（完全不会）至 3（非常积极）]

- 您对用于检测决策过程中的歧视行为的公平性监测系统了解多少？ [-3（完全不了解）至 3（非常了解）]
- 这是一道注意力检查问题。请从以下选项中找出并选择数字三（3）。 [-3 至3]

- 您对自动化招聘系统在招聘流程中的作用了解多少？例如，它如何处理求职申请？ [-3（完全不了解）至 3（非常了解）]
- 如果您捐赠自己的个人数据，了解这些数据将如何被使用，对您来说有多重要？ [-3（完全不重要）至 3（极其重要）]
- 如果您捐赠自己的个人数据，了解这些数据将如何受到保护，对您来说有多重要？ [-3（完全不重要）至 3（极其重要）]
- 您认为当前的自动化招聘系统在招聘过程中处理求职申请时有多公平？ [-3（完全不公平）至 3（极其公平）]
- 您有多愿意捐赠个人数据用于研究或其他用途？ [-3（完全不愿意）至 3（非常愿意）]

- 您认为当前的公平性监测系统在检测和减少决策过程中的偏见与歧视方面有多有效？ [-3（完全无效）至 3（非常有效）]

- 您认为捐赠数据会给您的隐私带来多大风险？ [-3（完全没有风险）至 3（风险极高）]

B.5 联合选择任务

通过五项联合分析任务表达他们对协议设计的偏好，图1展示了一个示例。

B.6 排序问题

拖动并排列以下七项属性：公平性目标、金钱激励、收集的信息类型、隐私保护机制、数据存储、监测主体和数据使用。

B.7 研究维度各选项接受度网格问题

接下来的 7 个问题中，您将对研究维度中的每个选项的接受程度进行评分。评分为 **-3** 表示最不偏好的选项，而 **+3** 表示最偏好的选项。

- **数据捐赠如何帮助改进自动化招聘系统?** :
  – 选出最合格的候选人 [-3 至 3]– 以相似方式对待不同人口统计群体 [-3 至 3]– 以相似方式对待相似个体 [-3 至 3]
- **谁在收集数据?** :
  – 审计机构和监管机构 [-3 至 3]– 开发自动化招聘系统的公司 [-3 至 3]
- **捐赠数据如何使用?** :– 评估自动化招聘系统 [-3 至 3]– 开发并评估自动化招聘系统 [-3 至 3]

- **如何补偿数据捐赠者?** : – 不提供金钱补偿 [-3 至 3]– 价值 1 美元的礼品卡 [-3 至 3]– 价值 5 美元的礼品卡 [-3 至 3]
- **收集哪些数据?** : – 个人数据 [-3 至 3]– 个人数据和招聘决策 [-3 至 3]– 个人数据、招聘决策和资历 [-3 至 3]

- **如何保护数据?** – 匿名化和加密 [-3 至 3]– 匿名化、加密和分布式存储 [-3 至 3]

- **谁负责将加密数据分开存储?**
  – 专业商业公司 [-3 至 3]– 研究中心 [-3 至 3]– 非政府非营利组织 [-3 至 3]

B.8 开放式问题

您是否想分享对调查内容、设计或其他相关方面的任何反馈或意见?