arXiv:2303.00738v1 [cs.CR] 1 Mar 2023

概率有多大？解释差分隐私中的 Epsilon 参数

Priyanka Nanayakkara[*†]
西北大学
priyankan@u.northwestern.edu

Mary Anne Smart[*]加利福尼亚大学圣迭戈分校 msmart@ucsd.edu

Rachel Cummings[†‡]哥伦比亚大学
rac2239@columbia.edu

Gabriel Kaptchuk[‡]波士顿大学
kaptchuk@bu.edu

Elissa Redmiles[‡]
马克斯·普朗克软件系统研究所
eredmiles@gmail.com

2023年3月2日

摘要

差分隐私（DP）是一种数学隐私概念，正日益应用于政府和业界。在差分隐私中，隐私保护具有概率性：其范围受隐私预算参数 ϵ 限定。健康和计算科学领域的既有研究发现，人们难以理解概率风险。然而，向贡献数据的人说明 ϵ 的影响，对于避免隐私作秀——将毫无意义的隐私保护包装成有意义的保护——并帮助人们做出更明智的数据共享决策至关重要。借鉴风险沟通和可用性方面的最佳实践，我们开发了三种向终端用户传达概率差分隐私保证的方法：其中两种说明赔率，另一种则提供差分隐私输出的具体示例。

我们通过一项情境问卷研究（$n = 963$），从三个指标对这些解释方法进行定量评估：客观风险理解、对差分隐私保证的主观隐私理解，以及自我效能。我们发现，基于赔率的解释方法比（1）基于输出的解释方法和（2）对 ϵ 信息一笔带过的最先进方法更有效。此外，与看到最先进的差分隐私解释相比，受访者在获知 ϵ 后更愿意分享数据；这种分享意愿会受到 ϵ 值的影响：隐私保护越弱，受访者分享数据的可能性就越低。

关键词 差分隐私 · 可用安全与隐私 · 隐私预算

1 引言

差分隐私（DP） [20] 是一种隐私的形式化定义，已应用于多个备受瞩目的数据分析流程，包括 2020 年美国人口普查的数据产品 [1] 以及谷歌 [23], Apple [4],微软 [18],和 Uber [76]等公司的内部指标测量工具。

随着差分隐私越来越多地应用于保护人们的隐私，部署差分隐私的组织必须有效传达那些决定系统隐私保护强度的实现细节所带来的隐私影响，这一点至关重要。缺乏这样的透明度，组织就有陷入“隐私作秀”的风险，[19, 71, 72] 这可能会让人们误以为自己受到了充分保护 [14, 75]。

[*]作者在哥伦比亚大学访问期间完成了本研究的部分工作。
[†]作者在加州大学伯克利分校西蒙斯计算理论研究所访问期间完成了本研究的部分工作。
[‡]作者对本研究的指导工作贡献相同。

尽管差分隐私提供了一个精确的框架来衡量最坏情况下的隐私损失，但研究发现，非专业人士难以准确评估差分隐私在现实世界中能够提供的隐私保护 [14, 80]。造成困惑的一个原因是，差分隐私的隐私保护具有概率性（即非绝对性）。具体而言，差分隐私根据无量纲的隐私预算参数 ϵ来界定隐私损失。差分隐私算法会向数据或分析结果中注入经过校准的统计噪声，其大小与 ϵ 成反比（具体取决于差分隐私模型），这意味着 ϵ 值越高，隐私保护越弱。
正如先前关于健康风险沟通的社会科学研究所指出的，向终端用户（即贡献数据的人）解释概率系统是一项艰巨的任务 [42, 70]。解释概率性的隐私风险（例如差分隐私所带来的风险）同样很有挑战性，甚至可能更加困难，因为系统的概率特性源于一种复杂且明确以数学为基础的过程，而非群体层面行为的变化。此外，差分隐私机制所提供的隐私保护缺乏上下文，也就是说，这些机制不考虑数据集或分析所处的社会背景。然而，隐私研究学者认为，人们会结合具体情境来理解隐私 [58]。

尽管 ϵ 至关重要，许多已部署的差分隐私系统却只在技术文档中说明 ϵ，而面向公众的隐私保护相关信息则完全忽略了所选 ϵ 的影响 [14, 19]。这尤其成问题，因为实践中使用的 ϵ 值差异极大，因此差分隐私系统在现实中所能提供的隐私保护也大相径庭 [11, 16]。

关于解释差分隐私的先前研究，要么回避向终端用户解释 ϵ这一复杂任务 [14, 24, 44, 80]，要么侧重于解决 ϵ对差分隐私特定部署的影响 [9, 15, 71],，例如解释随机响应机制 [9]。因此，我们目前缺少能够涵盖 ϵ 、且适用于所有差分隐私机制的差分隐私解释。若无法获得此类解释，部署差分隐私系统的组织就只能自行编写针对特定部署的差分隐私说明，而这些说明很可能未经科学评估；或者冒着让用户无法做出充分知情决策的风险。

我们通过开发并评估差分隐私解释方法来填补这一空白，这些方法能够直接 说明 ϵ的影响。我们的研究成果是一套框架，用于向终端用户传达 ϵ；它具有很强的可移植性，因此能够适配多种部署场景。我们的解释方法不依赖于对机制数学细节的描述，而是着重说明用户可能面临的选择会带来哪些具体影响，例如是否分享自己的数据。这与此前关于差分隐私传播的研究在概念上有所不同；后者关注使用差分隐私而非运行相应的非隐私系统所带来的影响 [9, 71]，或利用即使个体选择不分享数据也可获得的信息来学习属性 [24, 79, 80]。我们还在二元计数查询场景中，测试了我们的一种解释方法实例，以此评估这些方法。在第 6 节中，我们将介绍如何调整这些方法，以适用于其他查询和更复杂的场景。

差分隐私系统的新解释。 我们借鉴风险沟通和可用性方面的最佳实践 [22, 27, 28, 31, 39, 43, 47, 70]，开发了旨在帮助人们快速、轻松地推理差分隐私下概率隐私保证的解释方法。

具体而言，我们为 ϵ设计了三种解释方法。第一种解释方法（ODDS-TEXT）借鉴风险沟通领域的最佳实践方法，以文字描述一个人决定共享其数据时发生信息泄露的赔率；这是一种经过改编的差分隐私“教科书式”理解 [21]，它比较应用于相邻数据库的差分隐私机制的输出，而每个数据库分别对应某人共享或不共享其数据的情形。第二种解释方法（ODDS-VIS）通过频数框架可视化呈现相同的信息，这种方式可能有助于数感较弱的人更准确地判断概率 [27]。第三种解释方法（SAMPLE REPORTS）借鉴了可用安全与隐私（S&P）领域关于提升用户对安全与隐私技术理解的既有研究 [31]，向人们展示差分隐私机制可能产生的若干输出，以便将其数据共享选择所带来的影响具体呈现出来。

我们使用三项指标评估这些解释方法的效果：（1）客观风险理解；（2）主观隐私理解；（3）自我效能（个人对自身决策能力的信念 [53]）。此外，我们还研究解释方法与（4）数据共享意愿之间的关系。

总之，我们希望回答以下研究问题（RQ）：

RQ1： 在风险沟通中，哪些做法最有助于传达差分隐私（DP）所提供的概率隐私保证？具体而言，哪些做法能够有效提升人们的

(a) 差分隐私保证的客观风险理解，(b) 差分隐私保证的主观隐私理解，(c) 对数据共享决策的自我效能？

研究问题2：我们开发的解释方法如何影响人们的数据共享决策 ns?

我们通过一项情境问卷研究（$n = 963$）回答研究问题：将差分隐私解释嵌入具体的信息共享情境中，并依据上述标准，对这些解释与彼此及多种对照解释进行评估。

研究结果摘要。 我们发现，与 SAMPLE REPORTS（呈现隐私机制的样本输出）相比，采用基于赔率的解释（ODDS-TEXT 或 ODDS-VIS）时，人们对差分隐私保护的客观风险理解（研究问题 1a）更好。

尽管我们在客观风险理解方面取得了上述发现，但我们的解释都未能显著改善人们的主观隐私理解 （研究问题 1b），也就是说，人们并没有觉得自己理解了隐私保护。
此外，为评估自我效能（研究问题 1c），我们询问受访者是否觉得自己（1）掌握了足够的信息来做出数据共享决策，以及（2）有信心做出该决策。与一种不包含有关 ϵ 信息的先进 [80] 解释相比，我们测试的基于赔率的解释增强了人们认为自己掌握了足够信息来做出数据共享决策的感受，这表明不掩盖差分隐私保护的概率特性是有益的。然而，我们用来解释 ϵ 的 SAMPLE REPORTS 方法却产生了相反的效果：与对情境进行非常简单明了的描述、且不包含任何概率性隐私保护相比，它降低了人们认为自己掌握了足够信息来做出数据共享决策的感受，这表明这种解释反而使受访者感到困惑。

有趣的是，与最先进的解释相比，我们没有发现证据表明我们的任何解释会显著影响人们对数据共享决策的信心。相反，我们发现，人们对数据共享后果的总体担忧与其信心显著相关。

最后，我们研究了我们的解释对人们数据共享意愿的影响（研究问题2）。研究结果表明，无论采用哪种解释方法，当人们看到的差分隐私解释包含有关ϵ 的信息（而非不包含此类信息）时，他们分享数据的可能性都会大得多。最后，当受访者获得有关 ϵ的信息时，他们的数据共享决策会因 ϵ的变化而有所不同，从实证角度验证了理论上的主张：数据共享意愿取决于隐私机制 [21]的强度。

2 背景

差分隐私。 差分隐私 [20] 是一种数学隐私定义，它从宏观层面确保，无论分析中是否包含任何特定个体的信息，分析结果都应相似。差分隐私机制会在数据分析过程的某个环节加入经过精心校准的随机噪声，以掩盖个体层面的细节，同时保持整体层面的准确性。加入的噪声过多，就会淹没数据中的信号，使分析结果毫无用处；加入的噪声过少，为个体提供的隐私保护可能就没有实际意义。隐私预算参数 ϵ控制着这种权衡；较小的隐私预算能提供更强的隐私保证。下面给出差分隐私的正式定义：

定义： [Dwork 等人 等人。 [20]] 如果随机算法 $\mathcal{A}$： $\mathcal{D} \to \mathcal{R}$ 对于任意一对数据库 D， $D' \in \mathcal{D}$，
它们最多相差一 条记录，且对于任意子集 $S \subseteq \mathcal{R}, Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} \cdot Pr[\mathcal{A}(D') \in S]$，均满足条件。

差分隐私的实现通常采用本地或中央模型。[4] 在中央模型中，可信数据管理者会存储收集到的数据，并在发布有关数据的统计数据、图表或其他聚合洞见时，酌情添加噪声。在本地模型中，数据发送给管理者之前，会先向每个个体的数据添加噪声。尽管已有研究探讨了如何向本地模型中的终端用户解释隐私预算，[9,15,71],我们的研究则着眼于更具挑战性的任务：在中央模型的背景下解释隐私预算。
实现差分隐私最简单的机制之一是拉普拉斯机制 [20, 21]。在该机制中，数据收集者通过添加从以零为中心、尺度参数为 $\frac{1}{\epsilon}$ 的拉普拉斯分布中采样得到的噪声，来发布简单计数查询的结果。最终结果服从 $Lap(\mu, b = \frac{1}{\epsilon})$ 分布，其中 μ 是添加噪声前计数查询的真实值。[5]

向终端用户传达差分隐私。已有研究开始探讨如何向普通人群传达差分隐私 [9, 14, 15, 24, 44, 71, 80, 81]。对于本地模型，Smart 等人 [71] 和 Bullek 等人 [9] 将隐私保护的强度解释为比特被“翻转”的概率。不过，这种解释方式

[4]不过，也存在其他模型 [17]。[5]请注意，拉普拉斯机制的标准表述侧重于噪声 的分布，而非输出。我们的表述在数学上等价，并且与我们在第 3.2.2 节中对拉普拉斯机制的使用一致。

并不适用于中央模型，因为噪声是在汇总层面而非个体层面添加的。隐喻为解释差分隐私提供了另一种方式。例如，Karegar 等人 [44] 用图像模糊这一隐喻来类比向收集的数据添加统计噪声。这类隐喻有助于人们理解隐私—准确性权衡的存在：增加注入的噪声会加强隐私保证，但会降低准确性。然而，这些隐喻无法解释特定 ϵ 设置所带来的影响，而这正是我们在本研究中要解决的难题。

Xiong 等人 [81] 研究了如何传达隐私预算对位置数据隐私性和准确性的影响。作者为本地模型、中心模型和洗牌模型开发了插图（见 [5]），展示添加的噪声量如何影响隐私性和准确性。他们使用热力图比较添加噪声前后所收集位置数据的准确性。他们注入正向噪声，而不是无偏噪声，以避免出现负计数这一可能使不熟悉差分隐私的人感到困惑的问题。在我们的研究中，我们则选择接受添加无偏噪声所产生的有时违反直觉的结果，并考察终端用户对这些结果的看法。

Franzen 等人 [24] 借鉴了定量风险传播领域的文献，以解释差分隐私所提供的保护。尽管定量风险传播形式有助于理解，但数理能力较低的个体难以理解这些解释。我们在调查问卷中加入了数理能力测量，以确定我们的解释是否同样会使数理能力较低的个体处于不利地位。Franzen 等人 [24] 与我们的工作之间一个重要的区别在于，我们各自呈现的比较概率（即个体分享数据时发生负面结果的概率）不同：Franzen 等人呈现的是不进行数据收集时发生负面结果的概率，而我们展示的是个体不分享其数据、但其他所有因素保持不变时发生负面结果的概率。由于人们很少有权立即停止整个数据收集过程，因此我们认为，有必要探索这一独立的决策情境，以便与人们在现实世界中所做的决策更加贴合。

支持围绕以下内容进行决策 ϵ。关于如何传达 ϵ 影响的研究，往往聚焦于设定隐私预算的数据管理者或分析师。例如，已有多个用于差分隐私的界面（DPComp [37], Overlook [77], PSI（Ψ） [25], Bittner 等人 [6], DPP [41], ViP [56]），用于呈现 ϵ 对准确性和／或风险的影响，以支持更明智地设定隐私预算。虽然这些工具面向数据分析师和管理者，但也适用于向非专家传达差分隐私，因为通常并不假定这些分析师／管理者具备差分隐私背景或专业知识。因此，这些工具必须充分阐释相关的差分隐私概念，以支持围绕隐私预算进行决策。与此同时，我们注意到，终端用户通常不负责设定或分配隐私预算，而是必须就个体数据共享作出决策。

Hsu 等人 [40] 提出了一个经济学框架，供考虑分享数据的人们权衡分享与不分享数据所带来的货币成本，例如将数据用于科学研究时的情况。Wood 等人 [79], 在一份差分隐私入门指南中解释 ϵ 时，同样将数据共享决策表述为最坏情况下的货币损失（例如保险费上涨），即人们在差分隐私下分享数据可能承担的损失。Heffetz 和 Ligett [38] 则向经济学家介绍了 ϵ，并以计算平均工资值为背景，主要关注准确性结果。最后，Lee 和 Clifton [46] 通过考虑一名潜在攻击者来建模披露风险：该攻击者看到差分隐私机制发布的结果后，会基于此对分析中包含哪些人的信息这一信念进行贝叶斯更新。我们的基于赔率的解释方法同样会对攻击者看到差分隐私输出后更新的信念进行建模。

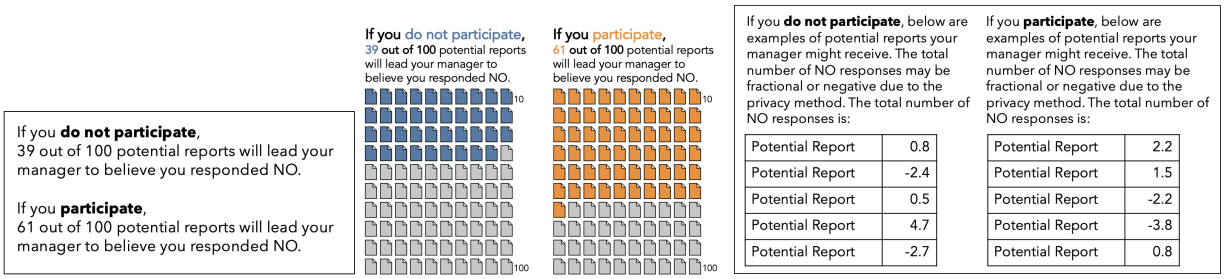
概率风险沟通。 许多研究已经确定了有效概率风险沟通的最佳实践，尤其是在医疗情境中 [22,47]。既有研究发现，将概率以频数形式呈现有诸多益处 [28,39]。概率风险沟通的一项挑战在于，人们常常会误解以比率表示的概率——例如，仅仅因为前一个比率使用的数字较小，人们可能会错误地认为概率为十分之一的事件不如概率为百分之十的事件可能发生 [3,83]。因此，在呈现比率以进行比较时，最佳实践是使用一致的分母 [47]。频数框架可视化（例如图标阵列）也可以补充数值风险沟通。与纯粹以数值呈现风险相比，图标阵列或许尤其能帮助数感较低的人提高理解程度 [27]。我们将这些研究发现融入对差分隐私的解释中，通过将概率以频数形式呈现并采用图标阵列来进行解释。

3 种解释方法 ϵ

我们介绍三种向终端用户解释 ϵ 的方法。这些方法适用于两种常见的数据共享情境：一种是提供数据可选，因此人们必须决定是否参与（或选择退出）；另一种是提供数据强制，因此人们必须决定是否如实回答（或不如实回答）。

借鉴健康风险沟通领域文献中的最佳实践 [22,39,47],，我们开发了两种解释方法（ODDS-TEXT 和 ODDS-VIS），重点是通过将负面事件发生的赔率 置于具体情境中来进行解释

隐私保障，即根据用户可以作出的决策可能产生的结果来阐释隐私保障。我们开发了第三种方法（样本报告），根据研究发现，具体示例有助于人们理解安全与隐私主题 [31]，该方法为受保护的数据提供具体示例。调查情境中实例化的每种解释方法（见第 3.2.1 节）示例均见图 1。



(a) ODDS-TEXT (b) ODDS-VIS (c) 样本报告

图 1：在可选设置和 $\epsilon = 0.5$下，将我们的解释方法应用于经理绩效评估场景的示例。

3.1 描述方法

ODDS-TEXT。 鉴于研究 [70] 发现，当风险以频数而非百分比的形式呈现时，人们能更有效地推理风险的赔率；这一发现也适用于隐私决策 [43]，尤其是差分隐私（DP） [24], 。因此，我们在 ODDS-TEXT 解释中将所有概率都表示为频数（图 1a）。具体而言，概率以“z 个潜在差分隐私输出中”的形式呈现，其中 z 为自然数；“差分隐私输出”可根据具体情境进行定制（例如，如果差分隐私输出被发布在报告中，解释可以改为“个潜在报告”）。具体而言，这种解释方法包含两个概率，分别对应于对手 $\mathcal{A}$基于先验知识与差分隐私输出，认为数据共享者提供了与实际值 d_{true} 相对应的信息的可能性：数据共享者参与与否，或如实与否地作出回应。数据共享者能够立即自主决定是否采取这些行动，因此，呈现与这些行动相对应的概率，与其数据共享决策过程直接相关。具体而言，ODDS-TEXT 解释采用以下形式：

> 如果你 *[不参与/不回应 d_{false}]*, x 在 100 个可能的 *[差分隐私输出]* 中，会有 $\mathcal{A}$ 个让人相信你回应了 d_{true}。如果你 *[参与/回应 d_{true}]*, y 在 100 个可能的 *[差分隐私输出]* 中，会有 $\mathcal{A}$ 个让人相信你回应了 d_{true}。

ODDS-VIS。 研究发现，图标阵列——一种频数框架可视化方法（图 1b）——有时能帮助数感较弱的人更准确地估计风险降低 [27]。因此，我们在 ODDS-TEXT 条件下的风险比文本描述中加入图标阵列。我们使用图标阵列，帮助人们直观地理解有许多可能的差分隐私输出，但 $\mathcal{A}$ 只能收到一个。图标的形状可根据具体情境进行调整。我们从上到下填充图标，因为研究表明，这种排列方式最有助于人们准确判断概率 [82]。图标颜色采用 Tableau 10 调色板 [73], ，该调色板在设计时考虑了常见的色觉缺陷。

样本报告。 借鉴 S&P 领域的既有研究，这些研究表明，具体示例能够提升用户对增强隐私的技术以及安全行为的理解 [31], ，样本报告方法展示了在数据共享者如实参与/作答（即共享 d_{true}）和不参与/不如实作答（即不提供数据或共享 d_{false}）这两种情况下，分析可能产生的五种差分隐私输出（图 1c）。同时呈现这两组可能的输出，使数据共享者能够比较其决策会如何影响这些数值，并了解其调查答复在多大程度上会左右差分隐私输出。

3.2 解释的情境化

为了运用上述描述方法，我们需要以下内容：(1) 一个具体场景，使数据共享者能够据此思考如何解释；(2) 上文所述参数的取值（例如，两种基于赔率的方法所需的概率）。本节将介绍一个假设的数据共享场景，并

$\epsilon = 0.1$	$\epsilon = 0.5$	$\epsilon = 2$	$\epsilon = 4$
$x = 48$	$x = 39$	$x = 18$	$x = 7$
$y = 52$	$y = 61$	$y = 82$	$y = 93$

表 1：基于赔率的解释方法所用数值。

概述如何针对给定场景恰当地计算解释所需的数值。我们在面向受访者的调查问卷中使用这一假设场景（更多详情见第 4 节）。如下文及第 6 节所述，我们的方法以及每种方法的计算技巧都可以扩展到许多其他差分隐私应用。

3.2.1 工作场所评估场景

想象一位员工（数据共享者）被要求分享对其经理的评估（$\mathcal{A}$），但担心如果经理认为员工给出了负面评价，就会对其进行报复。所有向这位经理汇报工作的员工都被要求分享评估，具体需要回答以下“是/否”问题：
你觉得经理给予了你足够的支持吗？这一场景描述的是在中心化差分隐私模型下进行的数据收集和分析：公司会收集未经加噪的回答，并使用拉普拉斯机制和特定的 ϵ，生成一份报告，其中包含计算得出的否定回答总数。不过，报告不会包含团队成员的姓名及其回答。请注意，只需在后续分析中代入其他噪声分布，我们的解释方法就能用于说明其他各种差分隐私机制（例如高斯机制）的隐私强度。

由于差分隐私保证通常是从最坏情况的角度来阐述的，我们进一步设计了这一情境，以呈现这样一种最坏情况：经理事先知道其他所有队友会如何回应。具体而言，我们假设他们都会回答“是”，而员工想回答“否”（d_{true}），因此员工可能会在汇总统计中被单独识别出来。可以根据其他情境调整队友的回答，并相应修改以下分析中的 μ。我们还假设，经理先验认为员工回答“否”的概率为 50%。有关如何更新 ODDS-TEXT 和 ODDS-VIS 的具体分析以适应非均匀先验，请参阅附录 E。

3.2.2 计算用于解释的 ϵ 值

下面介绍如何在上述工作场所数据共享情境中计算解释所需的值。无论员工处于强制数据共享设置还是可选数据共享设置，我们的解释方法计算出的值都相同。[6]

赔率-文本 和 赔率-可视化。 我们通过对经理的猜测过程进行建模，计算 ODDS-TEXT 和 ODDS-VIS 解释的数值。我们假设经理正确地认为，所有其他队友都会回答“是”，这与第 3.2.1 节所述的情景细节一致。因此，否定回答的实际数量为 0 或 1，具体取决于员工是否参与/如实回答。表 1 展示了多个隐私预算下与这一情景相符的数值。
具体而言，我们将经理的推断过程建模为：根据观察到的差分隐私输出，估计员工最有可能给出的调查答复。利用对差分隐私机制分布的了解，经理比较给定差分隐私输出时员工回答的后验概率，以确定最大似然估计（MLE）。设差分隐私机制的输出为 r。经理比较 $\Pr[E = r \mid E \sim Lap(\mu = 1,\ b = \frac{1}{\epsilon})]$ 与 $\Pr[E = r \mid E \sim Lap(\mu = 0,\ b = \frac{1}{\epsilon})]$,，分别对应员工如实报告/参与以及员工撒谎/不参与的情况，然后猜测后验概率较高的那个行为。我们计算阈值 $r_{\text{threshold}}$，使得 $\Pr[E = r_{\text{threshold}} \mid E \sim Lap(\mu = 1,\ b = \frac{1}{\epsilon})] = \Pr[E = r_{\text{threshold}} \mid E \sim Lap(\mu = 0,\ b = \frac{1}{\epsilon})]$，并报告 $\Pr[r < r_{\text{threshold}}]$（即表 1 中的 $\frac{x}{100}$）和 $\Pr[r > r_{\text{threshold}}]$（即表 1 中的 $\frac{y}{100}$）。

样本报告。 在我们的情境中，经理将收到一份报告，其中包含 否响应的总数。为了获取给定 样本报告 下的潜在输出 ϵ，我们从以下各项中分别随机抽取五次：Lap（$\mu = 1$、$b = \frac{1}{\epsilon}$），这些值对应于在共享 d_{true} 时可能产生的差分隐私输出；以及 Lap（$\mu = 0$、$b = \frac{1}{\epsilon}$），这些值对应于在没有参与/共享 d_{false} 时可能产生的差分隐私输出。我们不会对抽样值进行后处理，这意味着这些值可能是小数或负数。因此，我们在这些值之前加入一段说明：“由于隐私方法，否响应的总数可能是小数或负数。”

[6]在我们的 情形中，这些值相同，因为我们考虑的是计数查询，因此 敏感度不会改变。

4 评估

我们通过开展在线情境调查（$n = 963$）来评估我们的解释方法；这些调查旨在模拟现实世界中的决策行为 [30], , 具体而言，我们向受访者展示第 3.2.1 节所述的调查情境，以及使用我们的三种方法之一创建的 ϵ 解释，详见第 3.1 节。我们所在机构的机构审查委员会已批准这项研究。

4.1 调查情境

调查受访者被告知设想自己就是第 3.2.1 节中描述的数据共享者/员工，并处于可选设置或强制设置中。在可选设置中看到情境的受访者会被告知，虽然参与其所在公司的调查是可选的，但参与就意味着他们必然会如实作答（他们会回答“否”），也就是说，只要参与就不能撒谎。另一方面，看到强制设置的受访者会被明确告知，他们既可以如实作答（“否”），也可以不如实作答（“是”）。在两种设置下，情境细节都相同，只有措辞略有不同，用于表达公司调查是强制性的还是可选的。
遵循情境短文研究的最佳实践 [67], , 我们设计了这一情境，使其仿照工作场所和教育场景中普遍开展的绩效评估，从而增加受访者认为该情境可信且与自身相关的可能性。为确保所有受访者在理解我们对隐私保障的解释时，都基于相似的假设来认识自己为何希望在假设情境中保护其数据，我们明确说明了如果经理正确猜出某份负面调查答复，可能产生的负面后果（“如果你的经理认为你回答了否，他们可能会进行报复。例如，他们可能会给你负面的绩效考核、给你安排额外工作，或试图让你被解雇。”）。随后，我们要求受访者设想团队的绩效考核将通过一种“隐私保护方法”得到保护，并且为防止受访者搜索外部材料，我们在调查问卷的任何位置都不使用“差分隐私”这一术语。

随后，我们使用第 3 节所述的方法，采用四种隐私预算之一（$\epsilon \in \{0.1, 0.5, 2, 4\}$）来解释隐私保护机制的影响，这些预算代表了不同程度的隐私保护强度。ϵ [19]的设定没有统一标准，因此我们选择这些值，以体现学术差分隐私文献中通常推荐的范围（较小的ϵ 值，如 0.1），以及更符合现实部署情况的较大值 [16]。我们注意到，一些中心化差分隐私的现实应用所使用的隐私预算远高于我们的最大值（例如，美国人口普查局为 2020 年人口普查选区重划数据设定的总隐私预算为 19.61 [11]），但这些预算指的是经过多次查询累积的 ϵ，而我们的场景中只有一次查询。

在相同 ϵ 下收到样本报告的所有受访者，看到的随机抽样结果都相同。也就是说，为了确保受访者看到的内容一致，我们会预先生成这些随机抽样结果，而不会针对每位受访者动态生成新的结果。样本报告解释中显示的所有生成值都列于附录 C.3。[7] 两种基于赔率的解释中显示的值均见表 1。

我们通过认知访谈（$n = 12$）测试了调查问卷，遵循最佳实践 [65] , 并采用出声思维协议 [78], , 邀请潜在研究受访者参与，以进一步完善情境，使其清晰可信。根据这些访谈，我们反复修改了解释的引言，并进一步明确了信息披露可能造成的负面后果。此外，我们还邀请差分隐私、调查方法学和可视化领域的专家进行专家评审（$n > 10$），以测试调查问卷。

4.2 实验设计

我们采用被试间研究设计 $3 \times 4 \times 2$, 每位受访者在给定的情境类型（可选、强制）下，只会看到一种解释；该解释根据特定的解释方法（ODDS-TEXT、ODDS-VIS、样本报告）和隐私预算（$\epsilon \in \{.1, .5, 2, 4\}$）计算得出。我们还设置了两种对照解释，作为实验解释的比较对象：一种不提供任何隐私保护（确定性对照）；另一种则包含对差分隐私的高层次解释，但不提及 ϵ，该解释改编自 Xiong 等人 [80] （*Xiong* 等人对照）。

确定性对照：在我们的情境短文中，假设员工所面临的最坏情况是完全没有隐私保护。在这种情况下，情境中的负面后果风险是确定的（也就是说，受访者可以预期经理的判断结果是确定的）：如果受访者参与/回答“否”，其经理就会以 1 的概率认为他们回答了“否”；如果受访者采取相反做法，其经理就会以 0 的概率认为他们回答了“否”。这种“解释”不仅属于最坏情况，关于风险沟通的既有研究 [3, 7, 70, 83] 还表明，受访者最容易理解这种解释（客观风险理解），而且它可能会让受访者拥有

[7]为确保可复现性，我们使用日期作为机制的随机种子。抽取样本的代码可在此处获取：https://osf.io/ w59fv/?view_only=c42a3d68bf9d4f35abe488aab831e775.

更强的自我效能感，因为这种解释具有确定性，与我们的概率性解释形成对比。因此，我们采用这一确定性情境作为对照：为此，我们纳入与实验条件相同的情境文本，但省略对差分隐私的风格化描述以及对隐私保障的解释。

Xiong 等人对照组：我们还将实验中的解释与 Xiong 等人提出的当前最先进的差分隐私解释进行比较 [80]。由于这一解释没有提供关于 ϵ 的信息，因此有助于我们评估，增加有关 ϵ 的信息会如何影响人们对差分隐私保护的理解、数据共享决策中的自我效能，以及数据共享意愿。为了呈现这一对照组，我们采用与实验条件相同的情境文本，但将我们风格化的差分隐私描述和隐私解释替换为根据 Xiong 等人的内容改编的差分隐私描述 [80]（具体措辞见附录 C.2）。Xiong 等人提出了多种描述，我们采用了他们的“无名称差分隐私”描述，因为它最符合我们的情境；此外，他们的评估表明，人们认为这种描述易于理解，并且它有助于理解一道相关评估题目，该题目涉及第三方数据查看者。

4.2.1 评估指标与数据分享意愿

受访者回答了一些问题，以从三个指标评估我们的解释：(1) 客观风险理解，(2) 主观隐私理解，以及 (3) 自我效能。我们还研究了我们的解释与 (4) 受访者数据共享意愿之间的关系。

(1) 客观风险理解：我们设置了两道是非题，以评估这些解释是否能帮助人们理解情境中固有的风险：他们的经理是否会认为他们回答了“否”。除了“是”和“否”选项外，我们还提供了“我不知道”选项 [65]，以尽量减少偶然猜对的随机作答。此前关于向人们传达差分隐私的研究 [24, 71, 80] 也采用了类似的客观风险理解题。
在强制设置下，第一道题如下：

> 如果我在调查问卷中回答“否”（即如实作答），而不是“是”（即不如实作答），我的经理更可能认为我回答了“否”。

可选设置中显示的版本几乎完全相同，但询问的是：如果该人参与/不参与，经理会持有什么看法。

所有实验条件下的受访者以及确定性对照组的受访者都会回答这个问题。Xiong 等人对照组中的受访者不回答这个问题，因为该解释未说明 ϵ；因此不存在真实答案。[8] 对于我们的实验条件，在我们测试的所有隐私预算下，正确答案均为“正确”。对于确定性对照组，正确答案为“正确”。

第二个客观风险理解问题更具挑战性，但其形式与第一个问题相似（我们同样提供“正确”“错误”和“我不知道”三个选项）：

> *My m*经理认为我在调查问卷中回答 否 （即，回答)真的实可能性，是我回答 是 （即，不如实回答）时的两倍以上。

出于与第一道客观风险理解题相同的原因，只有处于实验条件和确定性对照组的受访者会看到这道题。对于我们的实验条件，$\epsilon \in \{2, 4\}$ 的正确答案是“正确”，$\epsilon \in \{0.1, 0.5\}$的正确答案是“错误”。对于确定性对照组，正确答案是“正确”。我们将得分设为 0–2 分，即答对的客观风险理解题总数。“不知道”视为答错。

(2) 主观隐私理解：我们请受访者使用 4 点语义量表（完全没有信心–非常有信心）评价自己对隐私保护的理解程度。先前向人们解释差分隐私的研究也曾询问他们对差分隐私的主观隐私理解 [9, 24, 71]。我们还请受访者用自己的话描述隐私保护，作为开放文本回答。只有实验条件下的受访者以及 Xiong 等人对照组的受访者会看到这些问题；接受确定性对照的受访者不会看到此问题，因为该对照并未描述隐私保护。

[8] 由于 Xiong 等人的对照组表明使用了差分隐私，因此可以认为第一道客观风险理解题有一个正确答案（“正确”）。然而，随着 ϵ 趋近于零，概率差异也趋近于零。因此，对于较大的ϵ，正确答案显然是“正确”；但对于非常接近 0 的较小 ϵ 值，概率之间在实际意义上可能几乎没有区别，而“错误”也近似正确。

(3) 自我效能：为了了解人们认为自己在多大程度上有能力根据我们提供的解释类型做出数据共享决策，我们向受访者提出三个有关决策信心的问题。首先，请他们使用 4 点语义量表评价自己是否有足够的信息来决定采取何种行动（类似于 Franzen 等人 [24] 归类为“主观理解”的问题）。其次，请他们通过开放文本回答描述，为了帮助自己做出决定，他们还希望获得哪些信息（如果有的话）。第三，请他们使用 4 点语义量表评价自己决定采取何种行动的信心。所有实验条件下的受访者都会看到这些问题。

（4）数据共享意愿：为了评估受访者的数据共享意愿，我们询问每位受访者是否愿意参与调查问卷（在自愿条件下），或是否会在调查问卷中如实作答（在强制条件下）。随后，我们请他们通过开放文本回答解释自己的理由。尽管受访者可以随时返回调查问卷之前的页面，但提交此题答案后，该答案便无法更改。所有受访者都会看到这些问题。

按照最佳实践 [65]，约一半受访者看到的 4 点语义量表会随机反向呈现。如果某位受访者看到的量表是反向呈现的，那么其调查问卷中所有对应的量表也都会反向呈现。

4.2.2 问卷结构

调查问卷受访者首先会看到说明，告知他们将阅读一个虚构情境并回答后续问题。接着，他们阅读情境的第一部分，其中介绍了一个关于其经理的假设性调查、他们希望如何回答“否”，以及这样做可能带来的后果。随后，我们通过请受访者使用 4 点语义量表（完全不担心—非常担心）评估自己的担忧程度，来判断他们是否确实担心这些后果；在第 5 节中，我们将其称为“基线关注度”。我们还设置了一道简单易答的理解检查题，并将两次尝试后仍未能正确作答的受访者筛除。

确定性对照组的受访者在此处结束情境。Xiong 等人对照组的受访者会阅读经改编的差分隐私解释。实验条件下的受访者则会看到以下对随机分布的抽象描述：

> 您的公司不会准确报告团队中有多少员工作出了否答。相反，公司会使用一种统计方法修改否答总数，从而生成许多可能的报告。因此，每份可能的报告所显示的否 答人数都可能略低于或高于实际人数。只有一份报告会被随机发送给您的经理。

随后，受访者会看到一段根据计算得出的解释。

随后，所有受访者都会回答一系列关于数据共享意愿、客观风险理解、主观隐私理解和自我效能的问题。最后，受访者还会回答有关数理能力 [48],互联网技能 [32], 和人口统计特征（性别、年龄、种族、教育程度、计算机科学/IT 教育或工作背景以及收入）的问题。最后一个问题为开放式文本题，用于机器人检测 [45]。受访者可以返回调查问卷的前几页查看信息，也可以通过提供的链接轻松访问包含情境和隐私保护完整说明的 PDF 文件。完整的调查问卷文本见附录 C.1。

4.3 参与者招募

参与者均通过 Prolific 招募，所在地为美国，且年龄均在 18 岁以上。我们进行了功效分析，以估算调查问卷所需的适当样本量 [12]；963 名受访者完成了在线调查问卷，另有 12 人完成了认知访谈。我们向参与认知访谈的参与者支付了 7.50 美元，访谈时长为 15 分钟（每小时 30 美元）。在各实验条件下，受访者获得 2.95 美元报酬（完成时间中位数：10.9 分钟；~每小时 16 美元），在各对照条件下则获得 2.30 美元报酬（完成时间中位数：9.2 分钟； ~每小时 15 美元）。受访者人口统计特征的详细分类见附录 A。

4.4 分析

我们的分析[9] 旨在研究：（1）我们的解释方法对若干因变量（DVs）的影响；以及（2）这些结果与社会人口统计属性之间的关系。既有研究发现，这些属性可能会影响隐私相关决策（例如，参见 [29, 33, 49, 52, 61–63, 66]）。为指导我们的分析，我们首先构建因果有向

[9]代码可在此处获取：https://osf.io/w59fv/?view_only=c42a3d68bf9d4f35abe488aab831e775。

无环图（DAG），这些图参考了既有研究，涉及人口统计属性与隐私顾虑（例如， [59]）、互联网技能（例如， [34,36]）以及数理能力（例如， [26]）之间的关系。这些有向无环图（DAG）见附录 D。

在主要分析中，我们构建了一组回归模型，用于研究自变量（IV）——解释方法（分类变量，详见下文）、情境设置（二元变量：可选或不可选）和隐私预算（ϵ 作为数值变量）[10]——对因变量的影响：客观风险理解、主观隐私理解、自我效能和数据共享意愿。对于二元因变量（数据共享意愿），我们构建逻辑回归模型；对于序数因变量（主观隐私理解和两项自我效能测量），我们构建序数回归模型；对于连续因变量（客观风险理解题的正确答题数量），我们构建线性回归模型。

构建回归模型时，我们将解释自变量视为分类变量。根据模型中的因变量，我们使用一种或两种对照解释作为比较的参考（基线）水平。正如第 4.2.1 节所述，并非所有因变量都适用于两种对照条件。在两种对照均适用的情况下，我们会构建多个模型，分别将每种对照设为解释自变量的参考水平。为了研究 ϵ 与自变量之间的关系，我们无法使用任何一种对照条件，因为目前没有用于呈现 ϵ信息的先进方法或对照方法。为了比较基于赔率的方法（ODDS-TEXT 和 ODDS-VIS）与基于具体示例、以可用性为导向的方法（样本报告）的效果，我们还针对每个因变量构建模型，并将样本报告设为解释的参考水平。

接下来，我们对自变量与年龄（数值型）、教育程度（分类变量）、性别（二元变量[11]）、基线关注度（序数型）、互联网技能（数值型）和数理能力（数值型）之间的关系进行次级分析。既有研究发现，这些因素可能会影响受访者解读数值相关信息的能力，例如我们在解释中呈现的信息 [24,70]。需要注意的是，根据关于情境短文调查，尤其是隐私情境短文的最佳实践指南，我们在本分析中纳入了基线关注度 [54]。我们拟合了仅包含人口统计特征（年龄、教育程度、性别）的模型，以及在调整上述人口统计特征后纳入基线关注度、互联网技能和数理能力的模型。

为支持统计分析的结果，我们对开放式文本回答进行了定性分析，以帮助 contextualize 突出的定量结果。两位作者审阅了约 10% 受访者的开放式文本回答子集，并共同编制了一份编码手册（见附录 B），用于记录能够阐释统计分析结果的回答主题。随后，其中一位作者又对约 20% 的回答子集进行了编码，以确保编码手册涵盖了大多数一般性观点；在第二轮编码过程中未发现其他编码。总计，我们对 283 份回答进行了编码。由于定性数据既未通过计数进行量化，也不是我们的主要研究重点，因此我们不报告编码者间信度 [55]。

4.5 局限性

我们的研究结果受到多个因素的限制，这些因素也普遍适用于大规模调查研究。首先，我们的样本可能未能涵盖具有代表性的群体。研究发现，针对隐私相关信念和感知的问题，Prolific具有相对较高的外部效度 [74]。然而，与众包研究的常见情况一样，我们的样本中年轻、受教育程度较高的个体占比较大，因此无法充分代表美国人口。我们的研究对象也仅限于居住在美国的人，这可能会限制研究结果在美国以外文化背景中的适用性。其次，我们通过多次认知访谈对调查情境进行完善，力求使其尽可能贴近现实且易于受访者理解，同时将调查问卷长度控制在合理范围内。然而，为了隐去某些细节，情境可能未能反映各种工作场所环境中的重要方面，而这些方面可能会影响人们对是否共享数据所作的决策。第三，尽管我们遵循最佳实践 [65] ，例如在适当情况下提供“我不知道”这一选项，以尽可能获取受访者的真实回答，但受访者的回答可能并不总是与他们的真实感受或在类似现实情境中会作出的决策一致。既有研究表明，情境短文研究能够有效帮助理解现实行为，尤其是在受访者投入度较高时。我们推测，受访者对这一虚构情境表达了高度担忧——超过75%的人表示“担忧”或“非常担忧”——这表明他们投入度较高。最后，我们的调查问卷只关注工作场所调查中的数据共享场景，因此研究结果未必能直接推广至其他场景。尽管我们是在单一情境中考察这些解释，但我们指出，为差分隐私开发的解释方法不受具体情境限制，因此未来研究可根据需要将这些解释方法应用于新的情境。

[10]我们在模型中纳入 ϵ ，而不是 e^{ϵ} ，因为 e^{ϵ} 会人为压缩相关隐私条件，尤其是在 ϵ [15]取值较小时。[11]我们向受访者提供了四个选项（男性、女性、非二元性别、自我描述），并允许他们选择多个 [69]。20 名受访者认同为非二元性别，另有一人认同为流动性别。我们的样本量太小，无法有意义地得出这些群体分别如何理解我们的解释。 因此，在建模时，我们将性别编码为某人是否认同为男性（无论其是否还认同其他性别）。

5 结果

基于调查结果，我们旨在评估这些解释的有效性（**研究问题1**），并研究这些解释如何影响人们的数据共享决策（**研究问题2**）。

5.1 解释的有效性（RQ1）

我们通过三个指标评估三种解释——ODDS-TEXT、ODDS-VIS 和样本报告——的有效性：(1) 对概率差分隐私保证的客观风险理解；(2) 对差分隐私保证的主观隐私理解；以及 (3) 自我效能。

5.1.1 客观风险理解（RQ1a）

我们按照第 4.4 节所述方法构建了一个线性回归模型（表2，左侧）[12]，其中因变量是每位受访者在客观风险理解题中答对的总数。我们发现，看到 ODDS-VIS 解释的受访者对隐私风险的客观理解高于看到确定性对照的受访者（$\beta = 0.28$，95% 置信区间 $= [0.10, 0.46]$）。另一方面，与看到确定性对照的受访者相比，看到样本报告的受访者在客观理解问题上的表现显著较差（$\beta = -0.43$，95% 置信区间 $= [-0.61, -0.25]$）。

我们还构建了一个因变量相同的线性回归模型，但将解释的参考水平设为 SAMPLE REPORTS 解释方法，这使我们能够将 ϵ 纳入模型（如第 4.4 节所述）。我们观察到，ϵ 增加与客观风险理解略有提升相关（$\beta = 0.03$，95% 置信区间 $= [0, 0.06]$）。随着 ϵ增加，我们提供的结果（即赔率或差分隐私样本输出）之间的差异也随之扩大；我们推测这使比较更容易。此外，与收到 SAMPLE REPORTS 解释方法的受访者相比，收到 ODDS-TEXT 和 ODDS-VIS 解释的受访者在这些问题上的表现显著更好（$\beta = 0.57$，95% 置信区间 $=[0.46, 0.68]$；$\beta = 0.71$，95% 置信区间 $= [0.60, 0.82]$）。

我们的次级分析（表 3）显示，在对年龄、性别和教育程度（以下统称“人口统计特征”）进行调整后，数理能力越强，客观理解水平越高（$\beta = 0.29$，95% 置信区间 $= [0.07, 0.52]$）；基线关注度最高者的客观理解水平也更高（$\beta = 0.27$，95% 置信区间 $= [0.06, 0.48]$）。我们推测，这可能是因为关注度较高的受访者会更努力地理解所呈现的信息 [10]。

变量	客观风险补偿 解释参考：确定性	客观风险补偿 解释性参考：样例报告	主观隐私理解 解释参考：Xiong 等人	主观隐私理解 解释性参考：样例报告
	β/CI	β/CI	OR/CI	OR/CI
解释：优势文本	.14 [−.04, .32]	**.57***** [.46, .68]	1.61 [.96, 2.68]	**1.68***** [1.23, 2.28]
说明：ODDS-VIS	**.28**** [.10, .46]	**.71***** [.60, .82]	1.46 [.88, 2.43]	**1.52**** [1.11, 2.06]
说明：样本报告	**−.43***** [−.61, −.25]	−	.97 [.58, 1.62]	−
设置：可选	−.05 [−.13, .04]	−.04 [−.13, .06]	1.14 [.9, 1.45]	1.19 [.92, 1.52]
ϵ	−	**.03*** [0, .06]	−	1.07 [.98, 1.16]

表2：左侧：考察正确回答的客观理解问题数量与实验自变量之间关系的线性回归模型结果。第一列中，解释自变量的参考水平为确定性对照；第二列中，参考水平为样本报告。请注意，第一列中的短横线（即 –）表示模型中未纳入 ϵ。第二列的样本报告行中有短横线，因为在该模型中，样本报告是参考水平。我们报告回归系数 β 以及这些系数的 95% 置信区间。$\beta > 0$ 表示增加，$\beta < 0$ 表示减少。右侧：考察主观隐私理解与左侧所列自变量之间关系的序数回归模型结果。解释的参考水平分别为 Xiong 等人的对照（第 3 列）和样本报告（第 4 列）。优势比 > 1 表示赔率增加，而优势比 < 1 表示赔率降低。* p< 0.05；** p< 0.01；*** p< 0.001。

[12]表格 that 包括 p值 are 可获取 位于: https://osf.io/w59fv/?view_only=c42a3d68bf9d4f35abe488aab831e775。

变量	客观风险补偿	主观隐私理解	*SE*（信息）	*SE*（信心）	分享意愿
	β/CI	OR/CI	OR/CI	OR/CI	OR/CI
Age	0 [−.01, 0]	**.99*** [.98, 1.00]	**.99**** [.98, .99]	**.99*** [.98, 1.00]	.99 [.98, 1.00]
教育程度：部分大学教育	0 [−.16, .15]	.85 [.58, 1.25]	1.05 [.73, 1.51]	1.17 [.81, 1.68]	.93 [.61, 1.43]
教育程度：学士 +	.12 [−.02, .26]	**.66*** [.46, .94]	.80 [.57, 1.13]	.94 [.67, 1.33]	.71 [.47, 1.05]
性别（男性）	−.02 [−.12, .08]	**1.65***** [1.29, 2.11]	**1.35*** [1.07, 1.70]	**1.32*** [1.05, 1.67]	.91 [.70, 1.19]
BC：有些担忧	.15 [−.08, .38]	.91 [.51, 1.62]	.75 [.42, 1.34]	**.44**** [.24, .80]	**.43*** [.18, .95]
BC：担忧	.18 [−.04, .40]	.94 [.55, 1.62]	**.56*** [.33, .98]	**.31***** [.17, .54]	**.31**** [.13, .66]
BC：非常担忧	**.27*** [.06, .48]	1.27 [.75, 2.13]	.80 [.47, 1.36]	**.39**** [.22, .69]	**.19***** [.08, .38]
互联网技能	−.02 [−.08, .04]	**1.39***** [1.19, 1.61]	**1.28***** [1.11, 1.47]	1.15 [1.00, 1.32]	.91 [.77, 1.08]
数理能力	**.29*** [.07, .52]	1.06 [.6, 1.87]	.82 [.48, 1.39]	.79 [.46, 1.36]	.59 [.31, 1.14]

表 3：次级分析结果：我们考察了因变量与人口统计特征（年龄、教育程度、性别（男性））、基线关注度（“BC”）、互联网技能和数理能力之间的关系。系数的解释见表2。

5.1.2 主观隐私理解（研究问题 1b）

接下来，我们比较参与者在看到我们的解释与 Xiong 等人的对照（表2，右侧）时对差分隐私保证的主观隐私理解，但未发现我们的任何解释方法相较于对照产生显著影响。我们构建了第二个模型，其中样本报告是解释方法的参考水平。在此模型中，我们发现，与样本报告相比，我们的 ODDS-TEXT 和 ODDS-VIS 解释方法与隐私保护理解感知的提升相关（优势比 $= 1.68$，95% 置信区间 $= [1.23, 2.28]$；优势比 $= 1.52$，95% 置信区间 $= [1.11, 2.06]$）。

在次级分析（表 3）中，我们发现，最高教育程度与较低的主观隐私理解相关（优势比 $= 0.66$，95% 置信区间 $= [0.46, 0.94]$），而认同自己为男性则与较高的主观理解相关（优势比 $= 1.65$，95% 置信区间 $= [1.29, 2.11]$）。年龄增长对主观理解降低也有轻微影响（优势比 $= 0.99$，95% 置信区间 $= [0.98, 1.00]$）。最后，在对人口统计特征进行调整后，我们发现，互联网技能提高与主观理解提升相关（优势比 $= 1.39$，95% 置信区间 $= [1.19, 1.61]$）。与先前研究一致，这些发现表明，多种社会因素会影响人们对自身隐私保障理解程度的看法。例如，老年人可能对自己了解安全与隐私主题的信心较低 [52]，男性可能在包括数字技能和使用在内的多个领域表现出更高的自信 [13, 35, 51, 60, 68]，而互联网技能较高的人可能认为自己对数字概念有更深入的理解 [35]。

5.1.3 自我效能（研究问题 1c）

我们从以下两个方面衡量自我效能：（1）受访者认为自己是否拥有足够的信息来做出这些决策；（2）他们在做出数据共享决策时是否感到自信（表 4）。

拥有足够的信息来做出决策。与确定性对照相比，我们发现，样本报告对受访者认为自己拥有足够信息来做出决策的感受具有显著负面影响（优势比 $= 0.48$，95% 置信区间 $= [0.30, 0.78]$）。我们推测，受访者认为样本报告中呈现的信息与他们所需的关键信息不一致。例如，一位受访者写道，他们“想知道 *[*报告中*]* 经理 *[*收到的数字*]* 是高于还是低于”否定回答的总数，这表明他们可能希望看到类似 ODDS-TEXT 或 ODDS-VIS 解释中的概率汇总。另一位受访者写道：“我想知道我的 *[*经理*]* 得到更高数字的可能性。”

与 Xiong 等人的对照组相比，收到 ODDS-TEXT 解释的受访者报告自己在我们的 4 点语义量表上，对是否拥有足够信息作出决定的感受高出一个等级的可能性高出 75% 以上（优势比 $= 1.76$，95% 置信区间 $= [1.07, 2.88]$）；收到 ODDS-VIS 解释的受访者这一可能性高出 65% 以上（优势比 $= 1.66$，95% 置信区间 $= [1.02, 2.72]$）。另一方面，SAMPLE REPORTS 未发现这种关联。

与 SAMPLE REPORTS 的受访者相比，我们发现，采用 ODDS-TEXT 和 ODDS-VIS 解释方法的参与者报告自己在是否拥有足够信息作出决定的量表上感受高出一个等级的可能性分别高出 65% 和 55% 以上（优势比 $= 1.66$，95% 置信区间 $= [1.22, 2.26]$；优势比 $= 1.57$，95% 置信区间 $= [1.16, 2.12]$）。我们未发现 ϵ 对是否拥有足够信息作出决定的感受有显著影响。最后，我们的次级分析（表 3）表明，年龄增长对是否拥有足够信息作出决定的感受有轻微的负向影响（优势比 $= 0.99$，95% 置信区间 $= [0.98, 0.99]$），而认同自己为男性则与较强的此类感受相关（优势比 $= 1.35$，95% 置信区间 $= [1.07, 1.70]$）。在调整人口统计特征后，互联网技能提高也与较强的此类感受相关（优势比 $= 1.28$，95% 置信区间 $= [1.11, 1.47]$）；基线关注度为“担忧”则与较弱的此类感受相关（优势比 $= 0.56$，95% 置信区间 $= [0.33, 0.98]$）。

决策信心。 与确定性对照或 Xiong 等人的对照相比，我们没有发现实验解释与决策信心之间存在显著关系。在第三个模型中，我们以样本报告作为解释参照，未发现 ϵ 对决策信心有显著影响。然而，与样本报告方法相比，我们的 ODDS-TEXT 和 ODDS-VIS 解释与决策信心提升相关（优势比 $= 1.58$，95% 置信区间 $= [1.16, 2.15]$；优势比 $= 1.50$，95% 置信区间 $= [1.11, 2.04]$）。通过次级分析（表 3），我们发现男性的决策信心更高（优势比 $= 1.32$，95% 置信区间 $= [1.05, 1.67]$）。我们还发现，三个更高水平的基线关注度（经人口统计特征调整后）都与较低的决策信心相关。我们认为，更担心负面后果的人可能会对该做出哪种数据共享决策感到更加矛盾。

变量	***SE***（信息） 解释参考：确定性	***SE***（信息） 解释参考：Xiong 等人	***SE***（信息） 解释参考：SR	***SE***（信心） 解释参考：确定性	***SE***（信心） 解释参考：Xiong 等人	***SE***（信心） 解释参考：SR
	OR/CI	OR/CI	OR/CI	OR/CI	OR/CI	OR/CI
解释：ODDS-TEXT	.80 [.49, 1.3]	**1.76*** [1.07, 2.88]	**1.66**** [1.22, 2.26]	1.55 [.94, 2.55]	1.19 [.73, 1.93]	**1.58**** [1.16, 2.15]
解释：ODDS-VIS	.76 [.47, 1.22]	**1.66*** [1.02, 2.72]	**1.57**** [1.16, 2.12]	1.47 [.89, 2.42]	1.13 [.70, 1.83]	**1.50**** [1.11, 2.04]
解释：样本报告	**.48**** [.30, .78]	1.07 [.65, 1.74]	–	.99 [.60, 1.62]	.75 [.47, 1.22]	–
设置：可选	.98 [.77, 1.25]	.99 [.78, 1.26]	1.00 [.78, 1.28]	1.10 [.87, 1.4]	1.09 [.86, 1.38]	1.10 [.86, 1.41]
ϵ	–	–	1.08 [1.00, 1.18]	–	–	1.07 [.99, 1.16]

表 4：考察“信息充足”（Info.）和“决策信心”（Conf）的序数回归模型。SR=样本报告；限定词=确定性对照。解释请参见表 2。

5.2 对数据共享的影响（研究问题2）

为回答研究问题2，我们考察实验中的解释在多大程度上影响人们的数据共享决策（表5）。与确定性对照和 Xiong 等人的对照相比，样本报告和 ODDS-TEXT 解释与数据共享意愿存在显著关联。与确定性对照相比，看到 ODDS-TEXT 解释的受访者分享数据的可能性约为其 2 倍（95% 置信区间 $= [1.14, 3.42]$）；看到样本报告的受访者分享数据的可能性约为其 3 倍（95% 置信区间 $= [1.73, 5.22]$）。考虑到样本报告解释似乎未能很好地帮助受访者理解客观风险，这一发现颇为有趣（见第5.1.1节）。与 Xiong 等人的对照相比，看到 ODDS-TEXT 解释的受访者分享数据的可能性超过 2 倍（95% 置信区间 $= [1.41, 4.3]$）；看到 ODDS-VIS 解释的受访者分享数据的可能性接近 2 倍（95% 置信区间 $= [1.12, 3.37]$）；看到样本报告解释的受访者分享数据的可能性接近 4 倍（95% 置信区间 $= [2.14, 6.57]$）。

我们还构建了一个逻辑回归模型，以考察因变量——数据共享意愿与 ϵ 之间的关系。我们发现，随着 ϵ 增加，隐私保护减弱，受访者分享数据的可能性会降低（优势比 $= 0.81$，95% 置信区间 $= [0.74, 0.89]$）。图2展示了在每种解释方法和 ϵ 值下表示愿意分享数据的受访者比例。此外，我们发现，与样本报告这一解释相比，受访者在收到 ODDS-TEXT 或 ODDS-VIS 解释时，分享数据的可能性更低（优势比 $= 0.65$，95% 置信区间 $= [0.45, 0.94]$；优势比 $= 0.51$，95% 置信区间 $= [0.35, 0.73]$）。我们推测，这可能与不同解释所带来的自我效能感差异有关（见第5.1.3节）。在以样本报告为参照的模型中，我们还发现，与在强制设置下如实作答相比，人们在可选设置下分享数据的可能性更低（优势比 $= 0.74$，95% 置信区间 $= [0.55, 0.99]$）——在强制设置下，人们可能会觉得

变量	分享意愿 解释参考：确定性	分享意愿 解释参考：Xiong 等人	分享意愿 解释性参考：样例报告
	OR/CI	OR/CI	OR/CI
解释：优势文本	**1.97*** [1.14, 3.42]	**2.45**** [1.41, 4.3]	**.65*** [.45, .94]
说明：ODDS-VIS	1.55 [.9, 2.68]	**1.93*** [1.12, 3.37]	**.51***** [.35, .73]
说明：样本报告	**2.99***** [1.73, 5.22]	**3.72***** [2.14, 6.57]	–
设置：可选	.80 [.61, 1.06]	.76 [.57, 1]	**.74*** [.55, .99]
ϵ	–	–	**.81***** [.74, .89]

表5：考察数据共享意愿与自变量之间关系的逻辑回归模型。关于表格的解读，请参见表2。

说谎让人不自在。当被问及解释自己的决策时，许多参与者明确表示希望诚实作答。例如，一位受访者写道：“无论如何都会说实话，因为 *[他们]* 无论结果如何都拒绝说谎。”最后，在调整人口统计特征后，较高的三个基线担忧水平对数据共享的影响减弱，这一点并不令人意外（表 3）。

开放式文本回答提供了进一步的背景信息对受访者决策的影响。许多受访者明确思考了自己的行为会如何改变经理相信自己回答了“否”的赔率。例如，一位受访者写道：“如果我参与调查，经理相信我回答了否的可能性只会略微高于我不参与的情况，所以我不妨发表自己的意见。”另一些人对隐私不太担忧，而是关注所收集数据的效用。例如，一位受访者认为，参与调查是“正确的做法，”，因为这可能改善同事的工作条件。

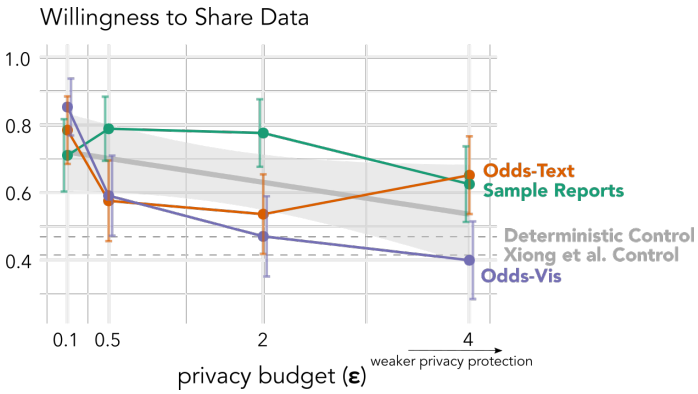


图2. 不同解释方法及 ϵ 下愿意共享数据的受访者比例，并附有 95% 二项分布置信区间。我们绘制了不同方法下的数据共享比例与 ϵ 之间的回归线（实线，灰色）。

6 讨论

我们发现，与基于示例的解释方法（SAMPLE REPORTS）相比，基于赔率的解释方法（在所有情况下均为 ODDS-VIS，在大多数情况下为 ODDS-TEXT）能够提升客观风险理解、主观隐私理解和自我效能。与现有的最先进差分隐私描述相比，基于赔率的方法也能增强人们认为自己拥有足够信息来做出隐私决策的感受；该描述 [80]，并未提供关于 ϵ 的信息。这表明，提供更多有关隐私保护的细节——即便是人们可能难以理解的概率信息 [70]——也能提升数据共享决策中的某些自我效能。此外，获得关于ϵ 信息的人，比获得未提及 ϵ的隐私保护描述的人更愿意共享数据。然而，与获得任一种基于赔率的解释的人相比，获得 SAMPLE REPORTS 的人更可能共享数据，尽管他们的客观风险理解相对较差。

最后，尽管受访者在获得隐私保护相关信息后会改变其数据共享决策这一点令人鼓舞，但我们也注意到，他们的数据共享决策、对这些决策风险的客观理解，以及作出这些决策时的信心，都取决于他们对其决策后果的担忧程度。特别担忧的人可能会投入更多认知资源，以准确推理概率隐私保证。[13]

传达效　　用影响。我们的工作建立在先前研究的基础上，这些研究表明，赔率能够有效地传达 - 隐私风险，并可用于纠正 与不保护隐私或不收集数据的替代方案进行比较 [24, 43]，以展示解释　　n

[13]参见 [2]，了解 可能影响数据共享决策的认知偏差　　ns.

传达赔率的解释方法可以帮助人们更好地理解概率性隐私保障。不过，我们的定性结果表明，在至少某些情境下，人们也可能会关心差分隐私对效用的影响。例如，一些受访者认为，准确传达他们的意见很重要，比如这样有助于改善工作场所环境。更高水平的差分隐私噪声虽然能提供更强的隐私保护，却会降低发布的统计数据的准确性和“效用”。乍看之下，准确性问题似乎更多属于数据管理者关注的范畴（例如第 2 节所述的界面），但我们认为，人们同样可能需要了解准确性的描述，以便有效判断其数据对后续用途的效用。这也有助于人们就数据共享作出知情决策。与先前研究的发现一致，即同时呈现准确性和隐私信息能够提升对人们在医疗场景中的数据共享决策的预测能力 [43]，也与旨在向人们传达差分隐私准确性影响的既有研究 [81], 相呼应，人们在差分隐私情境下进行判断时，除了考虑隐私，也可能受益于对准确性与隐私权衡的思考。因此，我们强调需要“超越”隐私，将效用视为尊重数据使用的关键组成部分 [64]。我们的结果揭示了如何</style)最有效地向人们传达效用影响。我们的 ODDS-TEXT 和 ODDS-VIS 解释方法展示了如何将 ϵ 映射为结果概率。解释还可以进一步提供从 ϵ到效用相关结果的映射。例如，如果收到一定数量的否定回答会要求经理完成额外培训，那么基于赔率的方法同样可以用来传达：在给定某人回答及特定 ϵ的情况下，达到该阈值的概率。这类信息有助于他们评估自己的回答能否带来实际结果，以及由此产生的效用。

迈向更复杂的场景。 我们的工作为向 ϵ 终端用户传达相关影响提供了一个框架；关于具体场景的细节可以进行调整，以便将我们的方法应用于差分隐私的各种实践。例如，我们没有明确建模对手对于某个个体以负面方式作出回应的先验信念，但我们的框架可以轻松应用于先验概率已给定、且对手执行贝叶斯更新的场景。详情请参见附录 E。若要将我们的基于赔率的解释方法应用于其他类型的查询（例如均值查询），我们建议基于假设检验框架 [50]计算概率。最后，作为未来的工作方向，我们认为，创建相关指南和工具具有很大潜力，使采用差分隐私的人员能够生成与其具体应用环境相一致的解释。这类工作将促进解释方法的实际应用，从而阐明 ϵ的影响。

围绕 ϵ的公众审议。 Dwork 等人 [19]提出，可以建立一个登记册，记录组织在应用差分隐私时所使用的 ϵ 值（以及其他实现细节）。他们认为，这样的登记册可以支持对不同差分隐私系统进行比较，并提高对差分隐私使用情况的问责，从而降低隐私作秀的可能性。我们的解释可以通过借鉴旨在向更广泛受众解释 ϵ 影响的工作，进一步提升此类登记册的作用，进而帮助促进围绕隐私预算的公众审议。更广泛地说，我们的解释方法以及其他解释 ϵ 的方法，可以通过让 ϵ 等实施决策更易于被不同背景和专业知识的相关方理解和讨论，来支持外部审计。例如，近期围绕美国人口普查局在2020年人口普查中使用差分隐私的讨论，展现了多个群体（计算机科学家、政策制定者、人口统计学家、非营利组织、公众等）之间有效讨论差分隐私相关方面所面临的挑战 [8, 57]。这些挑战部分源于差分隐私的技术细节（例如隐私预算）可能较为抽象或违反直觉，尤其是在计算机科学文献之外的语境中进行讨论时。面向广大群体、能够有效解释 ϵ 的内容，可以弥合沟通鸿沟，并促进公众围绕隐私政策展开审议。

7 致谢

我们感谢 Elizabeth Chase 和 Sean Kross 在分析方面提供指导，感谢 Jessica Hullman 审阅论文的早期草稿，感谢 Frauke Kreuter 对调查问卷设计提出反馈，也感谢 Chase Stokes 对可视化提出建议。此外，我们还感谢出席 2022 年差分隐私理论与实践（TPDP）研讨会和 2022 年情境完整性应用研讨会的与会者、西北大学 MU Collective 的成员，以及哥伦比亚大学隐私读书会的成员，感谢他们提出宝贵的讨论意见和反馈。所有作者均获得美国国防部高级研究计划局（合同编号 W911NF-21-1-0371）的支持。本材料中表达的任何观点、研究发现、结论或建议均属于作者个人意见，不一定反映美国政府或美国国防部高级研究计划局的看法。除美国国防部高级研究计划局的支持外，第三作者还获得了美国国家科学基金会资助项目 CNS-1942772（CAREER）、Mozilla 研究资助、摩根大通教职研究奖以及苹果隐私保护机器学习奖的部分支持。第四作者还获得了美国国家科学基金会向计算研究协会 CIFellows 项目提供的资助项目 #2030859 的支持；第五作者还获得了 Google 研究学者奖的支持。

参考文献

[1] John M. Abowd。《美国人口普查局采用差分隐私》。载于第24届 ACM SIGKDD知识发现与数据挖掘国际会议论文集，KDD '18，第2867页，美国纽约州纽约市，2018年。计算机协会。[2] Alessandro Acquisti、Idris Adjerid、Rebecca Balebako、Laura Brandimarte、Lorrie Faith Cranor、Saranga Koman- duri、Pedro Giovanni Leon、Norman Sadeh、Florian Schaub、Manya Sleeper、Yang Wang 和 Shomir Wilson。《隐私与安全助推：理解并协助用户在线选择》。ACM计算综述，50(3)，2017年8月。[3] Diego Alonso 和 Pablo Fernandez-Berrocal。《非理性决策：关注数字而非比率》。《人格与个体差异》，35(7)：1537–1547，2003年。[4] 苹果差分隐私团队。《大规模隐私保护学习》。2017年。[5] Andrea Bittau、Úlfar Erlingsson、Petros Maniatis、Ilya Mironov、Ananth Raghunathan、David Lie、Mitch Rudominer、Ushasree Kode、Julien Tinnes 和 Bernhard Seefeld。《Prochlo：人群分析的强隐私保护》。载于第26届操作系统原理研讨会论文集，第441–459页，2017年。[6] Daniel M Bittner、Alejandro E Brito、Mohsen Ghassemi、Shantanu Rane、Anand D Sarwate 和 Rebecca N Wright。《理解差分隐私在线主动学习中的隐私与效用权衡》。《隐私与保密性期刊》，10(2)，2020年。[7] Sidney T Bogardus Jr、Eric Holmboe 和 James F Jekel。《谈论医疗风险时的危险、陷阱与可能性》。《Jama》，281(11)：1037–1041，1999年。[8] boyd, danah 和 Sarathy, Jayshree。《差异化视角：围绕美国人口普查局使用差分隐私的认识论脱节》。《哈佛数据科学评论》（即将刊出），2022年。[9] Brooke Bullek、Stephanie Garboski、Darakhshan J. Mir 和 Evan M. Peck。《迈向理解差分隐私：人们何时信任随机响应技术？》载于2017年CHI计算系统中的人为因素会议论文集，CHI '17，第3833–3837页，美国科罗拉多州丹佛，2017年5月。计算机协会。[10] Richard L Celsi 和 Jerry C Olson。《涉入在注意与理解过程中的作用》。《消费者研究期刊》，15(2)：210–224，1988年。[11] 美国人口普查局。《人口普查局为保护2020年人口普查结果中的隐私设定关键参数》。2021年。[12] Stephane Champely。《pwr：功效分析的基本函数》，2020年。R软件包版本1.3-0。[13] Regina Ju-chun Chu。《家庭支持和互联网自我效能如何影响电子学习对老年人的作用——性别与年龄差异分析》。《计算机与教育》，55(1)：255–264，2010年。[14] 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 Giovanni Vigna 和 Elaine Shi （编），ACM CCS 2021，第 3037–3052 页。ACM Press，2021 年 11 月。[15] Inbal Dekel 、Rachel Cummings、Ori Heffetz 和 Katrina Ligett。《行为隐私弹性：概念化与应用》。https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4145498，2022年。[16] Damien Desfontaines。《差分隐私的现实应用列表》，2021年10月。Ted 正在写东西（个人博客）。[17] Damien Desfontaines 和 Balázs Pejó。《Sok：差分隐私》。《隐私增强技术论文集》，2020(2)：288–313，2020年。[18] Bolin Ding、Janardhan Kulkarni 和 Sergey Yekhanin。《私密地收集遥测数据》。《神经信息处理系统进展》，30，2017年。[19] Cynthia Dwork、Nitin Kohli 和 Deirdre Mulligan。《实践中的差分隐私：公开你的 epsilon参数！》。《隐私与保密性期刊》，9(2)，2019年。[20] Cynthia Dwork、Frank McSherry、Kobbi Nissim 和 Adam Smith。《在隐私数据分析中根据敏感度校准噪声》。载于 Shai Halevi 和 Tal Rabin（编），TCC 2006，《计算机科学讲义》第3876卷，第265–284页。施普林格，海德堡，2006年3月。[21] Cynthia Dwork 和 Aaron Roth。《差分隐私的算法基础》。《理论计算机科学中的基础与趋势》，9(3–4)：211–407，2014年。[22] Adrian Edwards、Glyn Elwyn 和 Al Mulley。《解释风险：将数值数据转化为有意义的图示》。《英国医学杂志》，324(7341)：827–830，2002年。

[23] Úlfar Erlingsson、Vasyl Pihur 和 Aleksandra Korolova。RAPPOR：随机化可聚合隐私保护序数响应。载于 Gail-Joon Ahn、Moti Yung 和 Ninghui Li（编），ACM CCS 2014，第 1054–1067 页。ACM Press，2014 年 11 月。[24] Daniel Franzen、Saskia Nuñez von Voigt、Peter Sörries、Florian Tschorsch 和 Claudia Müller-Birn。“我有隐私吗？如果有，隐私程度如何？”——运用风险沟通形式，让差分隐私更易于理解。arXiv 预印本 arXiv:2204.04061，2022 年。[25] Marco Gaboardi、James Honaker、Gary King、Jack Murtagh、Kobbi Nissim、Jonathan Ullman 和 Salil Vadhan。Psi (ψ)：隐私数据共享接口。arXiv 预印本 arXiv:1609.04340，2016 年。[26] Mirta Galesic 和 Rocio Garcia-Retamero。健康统计数感：基于概率抽样的跨文化比较。《内科学档案》，170(5):462–468，2010 年。[27] Mirta Galesic、Rocio Garcia-Retamero 和 Gerd Gigerenzer。利用图标阵列传达医疗风险：克服数感不足。《健康心理学》，28(2):210，2009 年。[28] Gerd Gigerenzer 和 Ulrich Hoffrage。如何在不加指导的情况下提高贝叶斯推理能力：频数格式。《心理学评论》，102(4):684，1995 年。[29] Hana Habib、Jessica Colnago、Vidya Gopalakrishnan、Sarah Pearman、Jeremy Thomas、Alessandro Acquisti、Nicolas Christin 和 Lorrie Faith Cranor。远离窥探的目光：分析隐私浏览的使用情况与理解程度。载于第十四届可用隐私与安全研讨会（SOUPS 2018），第 159–175 页，2018 年。[30] Jens Hainmueller、Dominik Hangartner 和 Teppei Yamamoto。通过现实世界行为验证情境短文与联合调查实验。《美国国家科学院院刊》，112(8):2395–2400，2015 年。[31] Marian Harbach、Markus Hettig、Susanne Weber 和 Matthew Smith。利用个人实例改进安全与隐私决策的风险沟通。载于 SIGCHI计算系统中的人为因素会议论文集，第 2647–2656 页，2014 年。[32] Eszter Hargittai。数字原住民？“网络一代”成员在互联网技能与使用方面的差异。《社会学探究》，80(1):92–113，2010 年。[33] Eszter Hargittai 等。Facebook 隐私设置：谁在意？《第一星期一》，2010 年。[34] Eszter Hargittai、Anne Marie Piper 和 Meredith Ringel Morris。从互联网接入到互联网技能：老年人中的数字不平等。《信息社会的普遍接入》，18(4):881–890，2019 年。[35] Eszter Hargittai 和 Steven Shafer。实际在线技能与感知在线技能的差异：性别的作用。《社会科学季刊》，87(2):432–448，2006 年。[36] Eszter Hargittai 和 Aaron Shaw。关注技能差距：互联网知识与性别在维基百科贡献差异中的作用。《信息、传播与社会》，18(4):424–442，2015 年。[37] Michael Hay、Ashwin Machanavajjhala、Gerome Miklau、Yan Chen、Dan Zhang 和 George Bissias。利用 DPComp 探索隐私-准确性权衡。载于 2016 年国际数据管理会议论文集，第 2101–2104 页，2016 年。[38] Ori Heffetz 和 Katrina Ligett。隐私与基于数据的研究。《经济展望杂志》，28(2):75–98，2014 年。[39] Ulrich Hoffrage 和 Gerd Gigerenzer。利用自然频数改进诊断推断。《学术医学》，73(5):538–540，1998 年。[40] Justin Hsu、Marco Gaboardi、Andreas Haeberlen、Sanjeev Khanna、Arjun Narayan、Benjamin C Pierce 和 Aaron Roth。差分隐私：一种选择 epsilon 的经济学方法。载于 2014 年 IEEE 第 27 届计算机安全基础研讨会，第 398–410 页。IEEE，2014 年。[41] Mark F St John、Grit Denker、Peeter Laud、Karsten Martiny、Alisa Pankova 和 Dusko Pavlovic。利用差分隐私为数据共享提供决策支持。载于 2021 年 IEEE 网络安全可视化研讨会（VizSec），第 26–35 页。IEEE，2021 年。[42] Daniel Kahneman、Stewart Paul Slovic、Paul Slovic 和 Amos Tversky。《不确定性下的判断：启发式与偏差》。剑桥大学出版社，1982 年。[43] Gabriel Kaptchuk、Daniel G Goldstein、Eszter Hargittai、Jake Hofman 和 Elissa M Redmiles。新冠肺炎应用程序做到什么程度才算足够好？效益、准确性与隐私对采用意愿的影响。arXiv 预印本 arXiv:2005.04343，2020 年。

[44] Farzaneh Karegar、Ala Sarah Alaqra 和 Simone Fischer-Hübner。探索适用于用户的差分隐私数据分析隐喻。载于第十八届可用隐私与安全研讨会（SOUPS 2022），第 175–193 页，2022 年。[45] Courtney Kennedy、Nick Hatley、Arnold Lau、Andrew Mercer、Scott Keeter、Joshua Ferno 和 Dorene Asare-Marfo。评估虚假受访者给在线民意调查带来的风险。皮尤研究中心，2020 年。[46] Jaewoo Lee 和 Chris Clifton。多少才够？为差分隐私选择 ε 。载于国际信息安全会议，第 325–340 页。施普林格，2011 年。[47] Isaac M Lipkus。传达健康风险的数字、文字和视觉格式：建议的最佳实践与未来建议。《医学决策》，27(5):696–713，2007 年。[48] Isaac M Lipkus、Greg Samsa 和 Barbara K Rimer。高学历样本在数理能力量表上的总体表现。《医学决策》，21(1):37–44，2001 年。[49] Eden Litt。理解社交网站用户的隐私工具使用情况。《人类行为中的计算机》，29(4):1649–1656，2013 年。[50] Changchang Liu、Xi He、Thee Chanyaswad、Shiqiang Wang 和 Prateek Mittal。从假设检验角度考察统计隐私框架。《隐私增强技术论文集》，2019(3):233–254，2019 年。[51] Mary A Lundeberg、Paul W Fox 和 Judith Punˊccohaˊr。信心十足却判断错误：置信度判断中的性别差异与相似之处。《教育心理学杂志》，86(1):114，1994 年。[52] Mary Madden。隐私、安全与数字不平等。数据与社会研究所，2017 年。[53] James E Maddux 和 Jennifer T Gosselin。自我效能。吉尔福德出版社，2012 年。[54] Kirsten Martin 和 Katie Shilton。结合情境看待移动应用隐私：一项关于用户对移动设备隐私期望的实证研究。《信息社会》，32(3):200–216，2016 年。[55] Nora McDonald、Sarita Schoenebeck 和 Andrea Forte。定性研究中的可靠性与评定者间信度：计算机支持的协同工作与人机交互实践的规范和指南。《ACM 人机交互论文集》，3(CSCW):1–23，2019 年。[56] Priyanka Nanayakkara、Johes Bater、Xi He、Jessica Hullman 和 Jennie Rogers。可视化差分隐私数据发布中的隐私—效用权衡。《隐私增强技术》（PETS），2022 年。[57] Priyanka Nanayakkara 和 Jessica Hullman。美国人口普查中围绕差分隐私的冲突因何而起？《IEEE安全与隐私》，(01):2–11，2022 年。[58] Helen Nissenbaum。《情境中的隐私》。载于《情境中的隐私》。斯坦福大学出版社，2009 年。[59] Dara O'Neil。互联网用户在线隐私顾虑程度分析。《社会科学计算机评论》，19(1):17–31，2001 年。[60] Frank Pajares。性别与自我调节学习中的感知自我效能。《理论与实践》，41(2):116–125，2002 年。[61] Yong Jin Park。男性和女性在隐私方面是否存在差异？互联网中的性别化隐私与（不）平等。《人类行为中的计算机》，50:252–258，2015 年。[62] Lee Rainie、Sara Kiesler、Ruogu Kang、Mary Madden、Maeve Duggan、Stephanie Brown 和 Laura Dabbish。在线匿名性、隐私与安全。皮尤研究中心，5，2013 年。[63] Elissa Redmiles。网络收益：美国社交媒体用户在社会资本、隐私保护和数字育儿实践方面的数字不平等。载于国际AAAI网络与社交媒体会议论文集，第 12 卷，2018 年。[64] Elissa M Redmiles。尊重性技术的必要性：超越隐私，第 307–316 页。施普林格，2021 年。[65] Elissa M Redmiles、Yasemin Acar、Sascha Fahl 和 Michelle L Mazurek。安全与隐私研究人员调查方法学最佳实践概要。技术报告，马里兰大学，2017 年。[66] Elissa M Redmiles、Sean Kross 和 Michelle L Mazurek。数字鸿沟在哪里？关于安全、隐私与社会经济状况的调查。载于 2017 年 CHI计算系统中的人为因素会议论文集，第 931–936 页，2017 年。[67] Amy Henderson Riley、Elizabeth Critchlow、Lyena Birkenstock、MariaLisa Itzoe、Katherine Senter、Nichole M Holmes 和 Steven Wesley Buffer。情境短文作为全球健康传播的研究工具：2000 至 2020 年文献系统综述。《医疗保健传播杂志》，14(4):283–292，2021 年。

[68] John A Ross、Garth Scott 和 Catherine D Bruce。《分数知识中的性别信心差距：学生信念与成绩关系中的性别差异》。《学校科学与数学》，112(5):278–288，2012。[69] Morgan Klaus Scheuerman、Katta Spiel、Oliver L Haimson、Foad Hamidi 和 Stacy M Branham。《促进性别平等与包容的人机交互指南》，2020。[70] Paul Ed Slovic。《风险感知》。 Earthscan 出版社，2000。[71] Mary Anne Smart、Dhruv Sood 和 Kristen Vaccaro。《借助差分隐私理解隐私作秀的风险》。《*ACM* 人机交互论文集》，6(CSCW2)，2022年11月。[72] Christopher Soghoian。《终结隐私作秀：揭露并遏制企业向政府披露用户数据》。《明尼苏达法律、科学与技术期刊》，12:191，2011。[73] Stone, Maureen。《我们如何设计 Tableau 10 的新配色方案》。*Tableau* 网站，2016。[74] Jenny Tang、Eleanor Birrell 和 Ada Lerner。《复现：我的研究结果如今能推广到多大范围？在线隐私与安全调查的外部效度》。载于第十八届可用隐私与安全研讨会（*SOUPS 2022*），第367–385页，2022。[75] Jun Tang、Aleksandra Korolova、Xiaolong Bai、Xueqiang Wang 和 Xiaofeng Wang。《Apple 在 macOS 10.12 中实现差分隐私所导致的隐私损失》。*arXiv* 预印本 *arXiv:1709.02753*，2017。[76] Katie Tezapsidis。《Uber 发布差分隐私开源项目》，2017。[77] Pratiksha Thaker、Mihai Budiu、Parikshit Gopalan、Udi Wieder 和 Matei Zaharia。《Overlook：面向大数据的差分隐私探索性可视化》。*arXiv* 预印本 *arXiv:2006.12018*，2020。[78] Gordon B Willis。《认知访谈实践：出声思维、口头追问及其他技巧，第42–63页。Sage 出版社，伦敦，2005。[79] Alexandra Wood、Micah Altman、Aaron Bembenek、Mark Bun、Marco Gaboardi、James Honaker、Kobbi Nissim、David R O'Brien、Thomas Steinke 和 Salil Vadhan。《差分隐私：面向非技术受众的入门指南》。《范德比尔特娱乐与技术法期刊》，21:209，2018。[80] Aiping Xiong、Tianhao Wang、Ninghui Li 和 Somesh Jha。《推动有效的差分隐私沟通，助力用户做出数据共享决策并提升理解》。载于*2020*年 *IEEE* 安全与隐私研讨会，第392–410页。IEEE 计算机学会出版社，2020年5月。[81] Aiping Xiong、Chuhao Wu、Tianhao Wang、Robert W Proctor、Jeremiah Blocki、Ninghui Li 和 Somesh Jha。《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。[82] Cindy Xiong、Ali Sarvghad、Daniel G Goldstein、Jake M Hofman 和 Çagatay Demiralp。《调查图标阵列中的感知偏差》。载于*CHI*计算系统中的人为因素会议，第1–12页，2022。[83] Kimihiko Yamagishi。《为何 12.86% 的死亡率比 24.14% 更危险：对风险沟通的启示》。《应用认知心理学：应用记忆与认知研究学会官方期刊》，11(6):495–506，1997。

人口统计特征

表 6 列出了调查受访者人口统计特征的详细情况。由于部分受访者没有回答人口统计问题，且部分受访者在种族/族裔/性别问题中选择了多个答案，因此各类别的人数相加后并不总是等于 963。另请注意，尽管只有五位受访者明确填写了“双种族”“多种族”或“混血”，但许多受访者在种族/族裔问题中选择了多个答案。受访者的年龄是用调查年份减去其提供的出生年份计算得出的。

人口统计属性	计数
性别	
流动性别	1
Man	479
非二元性别	24
女性	458
不愿回答	5
Age	
18至29岁	346
30至39岁	286
40至49岁	142
50+	173
不愿回答	16
种族/族裔	
西班牙裔或拉丁裔	104
黑人或非裔美国人	114
白人	718
美洲印第安人或阿拉斯加原住民	19
亚裔，夏威夷原住民 ，或太平洋岛民	84
混血 ，多种族 ，或双种族	5
独特的自由文本回答	3
不愿回答	9
教育程度	
高中或以下	140
部分大学教育	319
学士学位或以上	501
不愿回答	3
计算机科学/信息技术领域的教育/工作	
Yes	192
No	735
不愿回答	36
收入	
低于 10,000 美元	60
10,000 美元至低于 20,000 美元	64
$20,000 至不足 $30,000	89
$30,000 至不足 $40,000	95
$40,000 至不足 $50,000	79
$50,000 至不足 $65,000	131
$65,000 至不足 $80,000	103
$80,000 至不足 $100,000	93
100,000美元至不足125,000美元	77
125,000美元至不足150,000美元	49
150,000美元至不足200,000美元	38
200,000美元或以上	42
不愿回答	43

表 6：调查受访者人口统计特征

B 编码手册

表7展示了定性分析所用的编码手册。

Code	示例
对隐私保护的困惑／疑问	虽然有点难以理解，但我认为它会发送各种 报告否定回答的人数
回答效用	随机过程彻底掩盖了否定回答的真实数量 ；这对员工匿名性很有帮助，但对 经理来说却有些毫无用处。
赔率比较	如果我回答“否”，我的经理更有可能收到一份 报告，让他们相信我回答了“否”。
情境背景	还有谁会收到调查结果？人力资源部门会得到准确信息吗 因此，我相信如果老板报复我，人力资源部门会帮助我
诚实	我确实遇到过类似的情况，并因此遭到了报复 。尽管有过这样的经历，我仍然认为，对我来说，坦诚说明正在发生的事情很重要。 诚实地说明实际情况。

表7：编码手册

C 调查工具

C.1 情境文本

以下是强制设置下的情境文本：

设想你和另外四个人在同一个团队工作。你们五人都向同一位经理汇报。公司要求你们每个人都参加一项调查问卷。调查问卷中有以下是非题：
你觉得经理给予你的支持足够吗？
你曾与经理有过不愉快的经历，因此想回答“否”。
然而，你不想让经理知道你回答了“否”。如果经理认为你回答了“否”，可能会报复你。例如，他们可能会给你负面的绩效考核、给你安排额外工作，或试图让你被解雇。

你必须决定如何回答调查问题。你可以如实回答“否”，也可以不如实地回答“是”。

根据午餐交谈，经理显然知道，你所有其他队友都会回答“是”，表示他们觉得经理给予了足够的支持。另一方面，经理完全不知道你会如何回答。

经理会收到一份关于调查问卷结果的报告。这份报告会说明回答“否”的总人数。

尽管报告会对你们的姓名保密，但经理收到报告后，仍然可以据此猜测你的回答。

例如，假设报告显示有一个人回答了“否”。经理会认为那个人是你，因为经理相信其他队友都回答了“是”。

（确定性对照到此结束。）
（如果显示 *Xiong* 等人的控制项，它会出现在这里，之后场景结束。）

不过，你的公司会使用一种隐私保护方法，帮助防止经理正确猜出任何人的回答。

你的公司不会准确报告团队中有多少员工回答了否定回答。相反，他们会使用统计方法调整否定回答的总数，从而生成许多可能的报告。因此，每份可能的报告所显示的数字都可能略低于或高于否定回答的实际数量。

系统只会随机向你的经理发送一份报告。
请注意，由于采用了隐私方法，你的经理收到的具体报告可能会让其认为你回答了“否”，无论你在调查问卷中作何回答。

---

*(*赔率-文本、赔率-可视化或样本报告 会显示在此处。*)*

分享意愿：

- 根据上述情境和隐私保护的说明，你会对调查问题回答“否”（即如实作答）吗？（会/不会/不愿回答此问题。）
- 请简要说明你的推理过程。（开放文本回答）

客观风险理解：

- 如果我对调查问题回答“否”（即如实作答），经理相信我回答了“否”的可能性，会高于我回答“是”（即不如实作答）时。（正确/错误/我不知道/我不愿回答此问题。）
- 如果我对调查问题回答“否”（即如实作答），经理相信我回答了“否”的可能性会超过我回答“是”（即不如实作答）时的两倍。（正确/错误/我不知道/我不愿回答此问题。）

主观隐私理解：

- 根据场景和隐私保护的描述，你有多大把握理解了应用于调查结果的隐私保护？（完全没有把握/有一些把握/有把握/非常有把握/我不愿回答此问题。）

自我效能：

- 根据隐私保护的情境和描述，您有多大信心认为自己掌握了足够的信息，能够决定是否对调查问题回答“否”（即如实作答）？（完全没有信心/有一些信心/有信心/非常有信心/我不愿回答此问题。）
- 为了帮助您决定是否对调查问题回答“否”（即如实作答），您还希望获得哪些信息（如果有）？（开放文本回答）
- 根据隐私保护的情境和描述，您有多大信心决定是否对调查问题回答“否”（即如实作答）？（完全没有信心/有一些信心/有信心/非常有信心/我不愿回答此问题。）

C.2 Xiong 等人 [80]控制文本

不过，为了尊重您的个人信息隐私，分享给您经理的报告将包含采用隐私保护方法处理过的*“NO”*回复总数。该方法通过向聚合数据（例如*“NO”*回复总数）添加随机噪声来保护员工的隐私，使您的经理推断您在调查问卷中的回复的概率低于未采用隐私保护时的概率。

C.3 样本报告值

我们选定的四种隐私预算下，样本报告使用的数值：

	如果您不参与／回答“是”				
ϵ	#1	#2	#3	#4	#5
0.1	4.0	-12.2	2.4	23.7	-13.5
0.5	0.8	-2.4	0.5	4.7	-2.7
2	0.2	-0.6	0.1	1.2	-0.7
4	0.1	-0.3	0.1	0.6	-0.3
	如果您参与／回答“否”				
ϵ	#1	#2	#3	#4	#5
0.1	7.1	3.4	-14.8	-23.0	0.0
0.5	2.2	1.5	-2.2	-3.8	0.8
2	1.3	1.1	0.2	-0.2	1.0
4	1.2	1.1	0.6	0.4	1.0

D 有向无环图

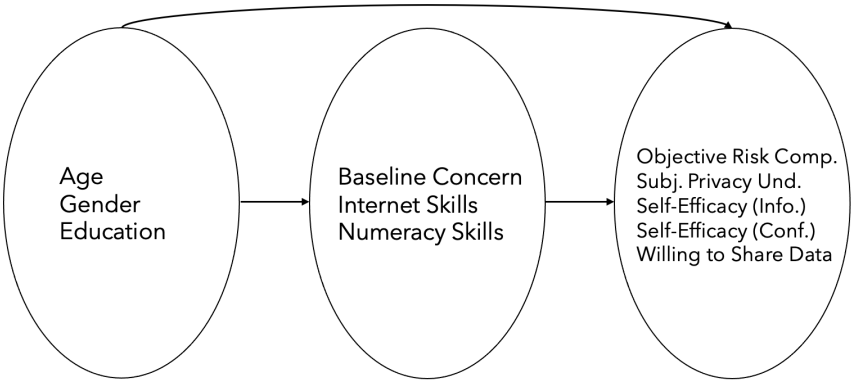


图3：该图展示了我们用于指导分析的有向无环图（DAG）。

E 建模对手的先验概率

为了让本文讨论的情景更加贴近现实且灵活，我们可以考虑这样一个事实：经理在看到隐私保护报告之前，可能已经对受访者会如何回答抱有先验信念。我们可以通过指定经理的先验信念，并假设经理会根据报告中的差分隐私输出进行贝叶斯更新，从而获得后验信念，来模拟经理的思考过程。我们将说明，在这一模型下，改变经理的先验概率，等同于移动差分隐私输出的阈值（$r_{\text{threshold}}$）：当差分隐私输出超过该阈值时，经理会认为一种结果（即受访者回答了 NO）比另一种结果（即受访者没有回答 NO）更有可能。

如果受访者回答 NO，经理会看到一个从以 1 为中心的拉普拉斯分布中抽取的样本。否则，经理会看到一个从以 0 为中心的拉普拉斯分布中抽取的样本。因此，经理的任务是猜测差分隐私输出 r 是从 $Lap(\mu = 1,\ b = \frac{1}{\epsilon})$ 还是 $Lap(\mu = 0,\ b = \frac{1}{\epsilon})$ 中抽取的。

我们使用贝叶斯定理，计算经理在查看差分隐私输出 r 后认为受访者回答“否”的可能性更大的概率（即经理的后验概率）。令 f_i 表示 $Lap(\mu = i,\ b = \frac{1}{\epsilon})$, 的概率密度函数，令 P_{no} 表示经理认为受访者回答“否”的先验信念。受访者回答“否”的更新概率（后验概率）为：

$$\frac{f_1(r)P_{no}}{f_1(r)P_{no} + f_0(r)(1 - P_{no})} = \frac{e^{-\epsilon|r-1|}P_{no}}{e^{-\epsilon|r-1|}P_{no} + e^{-\epsilon|r|}(1 - P_{no})}$$

同样，受访者没有回答“否”的更新概率为：

$$\frac{f_0(r)(1-P_{no})}{f_1(r)P_{no}+f_0(r)(1-P_{no})}=\frac{e^{-\epsilon|r|}(1-P_{no})}{e^{-\epsilon|r-1|}P_{no}+e^{-\epsilon|r|}(1-P_{no})}$$

我们可以通过求出使经理认为受访者回答“否”和未回答“否”的可能性相同的 r 值，来找到新的阈值 $r_{\text{threshold}}$。换句话说，我们令上述方程相等，并求解 r，得到：

$$r_{\text{threshold}}=\frac{\ln(1-P_{no})-\ln(P_{no})}{2\epsilon}+\frac{1}{2}$$

当 $\max\{\frac{1-P_{no}}{P_{no}},\frac{P_{no}}{1-P_{no}}\}\leq e^{\epsilon}$ 时（也就是说，只要先验概率不太极端）。

得出这个新阈值后，应用我们的解释方法就很简单了。我们只需根据个体可能做出的每种选择（例如，参与或不参与），计算差分隐私输出大于该阈值的概率。请注意，在我们的研究中，阈值设为 0.5。这是当 $P_{no}=0.5$ 时所得的 $r_{\text{threshold}}$ 值。