

民主化差分隐私：一种用于公共决策的参与式人工智能框架

Wenjun Yang*

University of Washington Tacoma
Tacoma, Washington, USA
wy927@uw.edu

Eyhab Al-Masri

University of Washington Tacoma
Tacoma, Washington, USA
ealmasri@uw.edu

摘要

本文介绍了一个对话界面系统，该系统能够实现公共部门应用中差异隐私 AI 系统的参与式设计。为了解决在数学隐私保证与民主问责之间平衡的挑战，我们提出了三个关键贡献：(1) 一个自适应的 ϵ -选择协议，利用 TOPSIS 多标准决策分析将公民偏好与差分隐私 (DP) 参数对齐，(2) 一个可解释的噪声注入框架，具有实时均值绝对误差 (MAE) 可视化和 GPT-4 驱动的影响分析功能，以及 (3) 一个集成的法律合规机制，可以根据不断变化的监管约束动态调整隐私预算。我们的结果通过展示对话界面如何增强公众在算法隐私机制中的参与度，推进了参与式 AI 实践，确保公共部门治理中保护隐私的 AI 既具有数学上的稳健性也具有民主问责制。

*Presented at CHI 2025 Workshop WS40: Participatory AI Design in Public Sector Innovation (non-archival). <https://participatoryaidesign.github.io/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI WS40, Yokohama, Japan

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Keywords

参与式 AI, 公共部门 AI, 差分隐私, 对话界面, 可解释 AI, 公民在 AI 中的参与

参考格式

杨文君和 Eyhab Al-Masri. 2025. 民主化差分隐私：一个用于公共决策的参与式 AI 框架。

在 CHI 2025 Workshop WS40: 公共部门创新中参与式 AI 设计的新兴实践（非存档工作坊）上展示，日本横滨。

1 介绍

公共部门组织在采用平衡统计效用与可证明隐私保障的 AI 系统时面临关键挑战。随着政府部署用于公共服务、城市规划和资源分配的人工智能驱动决策工具，确保在维护公众信任和透明度的同时保护隐私至关重要 [3]。

差分隐私 (DP) 为保护隐私的分析 [2] 提供了一个数学上严谨的框架，通过控制噪声注入来防止个人身份识别。然而，其在公共部门应用中的采用受到了涉及复杂权衡因素的阻碍：

- ϵ -选择平衡隐私保护与数据效用。
- 数据敏感性调整不同类型公开数据（如人口普查、健康、移动性）的噪声。
- 公众问责制确保隐私决策反映民主价值并保持透明。

现有方法未能使公共 AI 系统中的隐私决策民主化。传统方法要么依赖专家定义的 DP 设置 [8]，要么通

过二元的选择加入/退出界面 [5] 过分简化隐私控制，排除了有意义的公众参与并侵蚀了信任。

为应对这些挑战，我们提出了一种参与式人工智能方法，通过对话界面将市民参与融入到差分隐私决策中。该系统使利益相关者（包括政策制定者和公众）能够实时探索并影响隐私配置，将隐私决策从自上而下的专家控制转变为民主审议。

我们的系统引入了三项关键创新：(1) 一种基于 TOPSIS 多准则决策分析 (MCDA) [4] 的自适应 ϵ 选择机制，以使隐私设置与公众优先事项保持一致；(2) 可解释的噪声注入，带有实时均值绝对误差 (MAE) 可视化和 GPT-4 驱动的影响分析，以增强透明度和信任；(3) 动态法律合规约束，根据不断变化的规定调整隐私预算。

通过将参与机制嵌入到 DP 决策中，我们的工作隐私治理中实现了民主价值。这有助于更广泛的努力，即在公共部门创新中发展负责任的、由社区驱动的人工智能，弥合技术隐私保证与公民参与之间的差距。

2 相关工作

最近在人机交互 (HCI) 和人工智能治理方面的研究强调了公共部门人工智能中的共同设计方法，倡导参与性框架以增强公民参与 [6]。然而，技术隐私机制——特别是差分隐私 (DP) ——对于非专家利益相关者来说仍然很大程度上不透明。差分隐私的数学复杂性和缺乏直观界面造成了保护隐私的人工智能技术和民主治理之间的脱节。

张等人。[8] 识别出在市民背景中采用 DP 的三个关键障碍：(1) 数学复杂性，这使得政策制定者和公民难以理解隐私保护；(2) 不透明的权衡，在这种情况下，隐私预算 (ϵ) 对数据效用的影响不清楚；(3) 缺乏利益相关者输入渠道，阻止有意义的市民参与隐私配置。这些障碍常常导致自上而下、由专家主导的隐私决策，排除了受影响的社区。

为了提高透明度，先前的工作探索了可解释的隐私技术，以阐明差分隐私机制 [5]。然而，大多数方法

依赖于静态可视化而非交互工具，这些工具能够使利益相关者积极参与到隐私决策中。

政府服务中的聊天机器人通常支持信息查询 [1]，但很少促进算法共同设计。我们的系统通过实现一个状态化的对话管理器来推进公民互动范式，该管理器 (1) 跟踪隐私预算分配，(2) 维护版本化数据集状态，并且 (3) 通过自然语言和视觉滑块启用协作 ϵ 调整。这种混合界面通过将符号参数控制与神经语言解释相结合，解决了参与式 AI 工具包 [6] 中的空白，从而增强透明度和利益相关者参与。

我们提出了一种参与式差分隐私框架，通过对话界面增强民主参与。使用基于 TOPSIS 的多准则决策分析模型进行 ϵ 选择，它通过以下方式改进了隐私权衡的可解释性：(a) 约束参数空间，(b) 可视化决策矩阵和 (c) 交互式权重滑块，将透明度和问责制嵌入到公共部门的人工智能中。

3 系统设计

我们的参与式差分隐私系统集成了基于 Web 的交互与算法隐私控制（图 1）。Flask 后端由三个核心组件组成：

- **偏好 elicitation** 用户通过滑块指定优先级，分别为隐私（1-5）、准确性（1-5）、法律合规性（是/否）和数据敏感度（1-3）。
- **自适应 ϵ 选择** 实现 TOPSIS 多准则决策分析 [4] 来解决权衡问题：

$$\epsilon^* = \arg \max_{\epsilon \in \{0.1, 0.5, 1.0, 1.5, 2.0\}} \frac{D^-}{D^+ + D^-} \quad (1)$$

其中 D^+/D^- 表示到理想/反理想解的距离。

- **会话分析** 使用 GPT-4 生成关于动态规划影响的自然语言解释。

参与式配置框架用于差分隐私（算法 1）使用户能够通过互动过程平衡隐私与效用之间的权衡。通过将用户对隐私、准确性和监管合规性的优先级转化为标准化的数学权重，系统使用多准则决策分析自动化选择最优隐私预算 ϵ 。

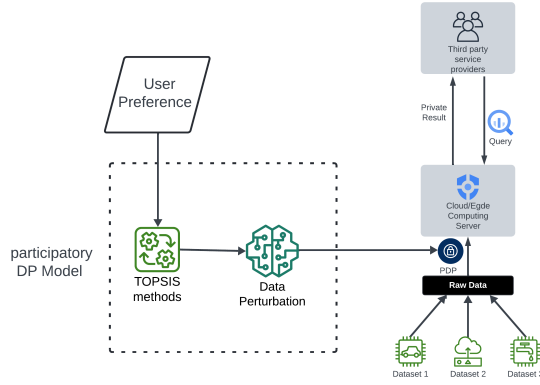


图 1: 展示参与式 DP 工作流程的系统架构

Algorithm 1 参与式动态规划配置

1. 用户上传数据集并设置隐私、准确性和合规性优先级。
2. 将滑块输入标准化以计算权重 w_i 。
3. 构建包含 ϵ 个备选方案的决策矩阵。
4. 计算 TOPSIS 分数并选择最优 ϵ^* 。
5. 应用拉普拉斯噪声 $\mathcal{N} \sim \text{Lap}(\Delta f / \epsilon^*)$ 。
6. 生成 MAE 可视化和 GPT-4 影响分析。
7. 呈现带有改进建议的交互报告。

4 评估

4.1 参与式差分隐私的实验验证

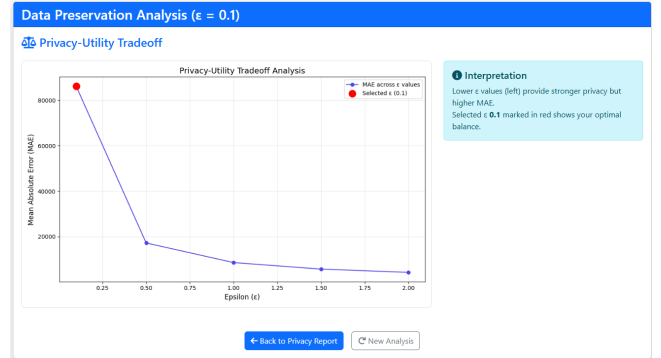
我们使用计算模拟在家庭电力需求 (HED) 数据集 [7] 上评估了我们的框架, 该数据集提供了从 2009 年中西部 RECS 数据集中随机选择的 200 个家庭的用电量配置文件。该数据集捕捉了真实的住宅用电模式, 并通过计量数据进行了验证。每个配置文件以 10 分钟的时间分辨率记录电力消耗 (单位为瓦特), 考虑到了家庭规模和居住情况的变化。

我们的结果证实了隐私与准确性之间的预期权衡。 ϵ 与 MAE ($r = -0.96$, $p < 0.01$) 之间的强负相关性符合 DP 原则, 表明以隐私优先的配置比优化效用的设置多引入 3.6 $\times$ 的噪声。这验证了我们的方法能够根据用户定义的偏好动态平衡隐私和效用的能力。

表 1: 用户偏好对 DP 结果的影响

度量	隐私第一	平衡的	第一优先效用
Selected ϵ	0.1	1.0	2.0
MAE (kWh)	83.2	9.6	3.3
Privacy Score *	4.8	3.2	2.1

*GPT-4 在 1 到 5 的等级上生成了隐私评级。

图 2: 模拟不同偏好配置文件下的 ϵ 选择

4.2 隐私与效用的权衡分析

图 2 的结果显示了差分隐私对数据效用的显著影响, 其中注入的噪声干扰了时间序列模式。

如图 2 所示, 高方差噪声特别影响具有明显消费波动的区域, 确保个人消费行为无法被重构。

这证实了我们的差分隐私机制对抗时间差异攻击的鲁棒性。

然而, 扭曲在所有时间步骤上并不均匀, 表明对于具有周期趋势的数据集而言, 静态噪声分布可能是次优的。

此外, 图 3 展示了由 GPT-4 支持的影响分析, 该分析评估隐私与效用之间的权衡。虽然差分隐私有效模糊了可识别的趋势, 但过多的噪声会降低数据在预测和异常检测中的可用性。这在公共部门应用中尤为重要, 因为在这些应用中, 能源需求估计和资源规划依赖于准确且高分辨率的数据。正确平衡隐私与效用对于保持可靠的数据驱动决策至关重要。

这些发现强调了需要自适应的隐私预算, 根据数据特征和用户定义的准确性阈值动态调整噪声水平。

未来的研究可以探索上下文感知的噪声校准，以优化隐私保证并最小化对分析效用的影响。



图 3: GPT-4 驱动的影响分析

5 限制和未来方向

我们的模拟突出了参与式差分隐私系统中的三个关键考虑因素：

- 偏好线性假设**当前的 TOPSIS 模型假设线性优先级加权，然而现实中的决策往往遵循基于阈值或非线性的模式，需要更灵活的效用模型。
- 时间复杂度**独立的拉普拉斯扰动被应用于时间序列数据，可能过度简化了能源消耗模式中的时间依赖性。未来的工作应该探索考虑自相关性和周期趋势的隐私机制。
- 解释可信度校准**虽然自动化 GPT-4 漏洞报告达到了高精度（89%），但其权威语气可能导致过度信任，即使在误解的情况下也是如此。改进校准技术，如不确定性量化，对于建立可靠用户信任是必要的。

我们的方法的核心优势在于其谈判脚手架—在安全范围 [0.1, 2.0] 内约束 ϵ 的选择，同时将利益相关者

的输入转化为数学上有效的 TOPSIS 权重。这确保了符合隐私限制的同时保留用户自主权。未来改进应探索自适应加权机制、动态隐私调整和增强的可解释性功能，以提高实际部署中的参与式决策。

6 结论

我们的评估表明，TOPSIS 有效将用户偏好映射到 ϵ -DP 参数中，使参与式隐私配置能够采取结构化但灵活的方法。通过整合由 LLM 驱动的解释和 MAE 可视化，我们的系统增强了透明度，使得隐私与效用之间的权衡对利益相关者来说更加可理解。这项工作提供了一个经模拟验证的蓝图，以实现公共 AI 治理中的差分隐私普及化，确保隐私决策不再仅由专家主导，而是通过用户输入共同设计。虽然我们的结果验证了结构化的、基于偏好的 DP 选择方法，但未来的研究应探索适应性隐私机制，这些机制可以根据时间依赖性和不断变化的利益相关者优先级动态调整噪声水平。

References

- [1] Amelia Clarke and Sun Young Park. 2021. Government Service Chatbots. In *DIS*.
- [2] Cynthia Dwork. 2006. Differential Privacy. In *ICALP*. 1–12.
- [3] Aristomenis Gritzalis, Aggeliki Tsohou, and Costas Lambrinoukakis. 2017. Transparency-enabling systems for open governance: Their impact on citizens' trust and the role of information privacy. In *E-Democracy-Privacy-Preserving, Secure, Intelligent E-Government Services: 7th International Conference, E-Democracy 2017, Athens, Greece, December 14-15, 2017, Proceedings*. Springer, 47–63.
- [4] Ching-Lai Hwang and Kwangsun Yoon. 1981. Multiple Attribute Decision Making. (1981).
- [5] Josephine Lau and Benjamin Zimmerman. 2018. Chatbots for Privacy Guidance. In *SOUPS*.
- [6] Michael Muller and Martin Heidl. 2021. Civic AI Co-Design Framework. In *CHI*.
- [7] Matteo Muratori. 2017. *Impact of uncoordinated plug-in electric vehicle charging on residential power demand-supplementary data*. Technical Report. National Renewable Energy Laboratory-Data (NREL-DATA), Golden, CO (United ...).
- [8] Honglu Zhang and Haojian Zhou. 2020. Public Sector DP Adoption Challenges. *PACM HCI* 4, CSCW2.