(*Work In Progress*) From Pink Tax to Blue Gambling: A novel harm-centric Framework to evaluate the impact of Ad Ecosystems.

Mathilde Raynal
EPFL

Theresa Stadler
SDSC

Carmela Troncoso
EPFL, MPI-SP

***Abstract*—Targeted advertising systems use personal data to deliver customized advertisements, raising significant concerns about discriminatory practices or overexposure to harmful content. One example of such practice is the Pink Tax, where users identified as women are subject to a price increase. Solutions in this space focus primarily on enabling advertising while ensuring privacy protection, overlooking the broader social consequences of algorithmic targeting on individuals and communities. In this talk, we introduce a framework for evaluating targeted advertising systems through *harm* caused by exposure to ads and *advertising utility* in terms of alignment of ads with user profile characteristics. Considering both dimensions jointly enables to understand the impact of ads on users while considering (potentially conflicting) interests of all parties involved. Our framework operates in a black-box manner: it feeds (simulated) user profiles to the ad ecosystem via browsing and collects the ads that users are served. Then, it analyzes the ads to infer harmful exposure; and uses an LLM-based framework to evaluate relevance of ads.**

We evaluate Google's current ad ecosystem. Using browsing data from 36 German users, we collect 704 personalized ads across categories such as gambling, fashion, technology, and finance. We evaluate these ads and find evidence of a gender bias in gambling advertisement targeting, with 77% of gambling ads delivered to male users, most of whom had not previously visited gambling-related websites.

1. Our Framework

We propose a harm-centric framework to evaluate ad ecosystems which consists of three steps: *profile creation*, *ad collection*, and *ad processing* (see Fig. 1).

Profile creation. We generate user profiles that can be used to trigger personalization of ads by the ad ecosystem under study. We construct profiles by reproducing the browsing of the pages associated with real users in an automated fashion and collect the resulting cookies. We use 36 browsing profiles from a dataset of real users' browsing behavior [1] that also contains self-reported demographic attributes: gender (binary) and age (buckets).

Ad collection. We gather ads selected and delivered by the ad ecosystem to our simulated users. Once we generate a profile, we visit a pre-selected set of 5 websites that are commonly visited by users and support Google Ads: sciencealert.com, genius.com, kicker.de, forbes.com, and merriam-webster.com. We extract the ads served to each users using adFisher [2] which enables us to automatically screenshot ads on the page. To ensure that we collect the same amount of ads per website for all users, we modify AdFisher to hard code the placements of the ads on the 5 websites we study. We choose 5 websites as a compromise between manual effort to hardcode and increasing the number and diversity of ads.

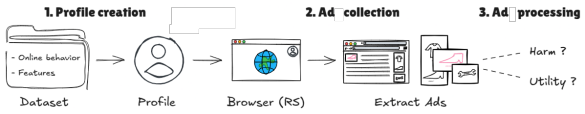

Figure 1: Overview of our approach

Ad processing. We evaluate the collected ads to compute *utility*, which captures whether the system fulfills its (economic) objective via measuring whether the ads are relevant to the user they are served to; and *harmfulness*, which captures whether the ads cause a harm to the user, and result on societal risks. This enables to measure the impact of profiling strategies on harm and business models.

2. Our Metrics

Utility. Typical notions of utility in advertising are largely consumption-oriented, in the sense that they focus with user engagement and/or purchase of the advertised product. Since many metrics are click-centric [3], we propose to use the *likelihood that a user clicks on an ad compared to a random baseline* as a metric for utility.

Since we do not have access to the participants that generated the profiles used in data collection, we simulate them leveraging TinyTroupe [4], a library for simulating LLM-driven personas with defined interests and behaviors. We choose TinyTroupe because advertising applications is one of its main proposed uses and it has been validated to approximate users' behavior. To maximize the quality of the simulation given our purpose, we select the best way to generate personas from the browsing history and demographic features and to perform screenshot-to-description conversion of ads, since the input is text only. Eventually, we generate a persona for every of the 36 users. We then present them ads and ask them their likelihood to click on it. We use the

agents' response as the utility of the ad towards this user. While LLM personas might not *exactly* reproduce the clicks of the users in our dataset, they provide a signal whether the ads align with interests and characteristics inferrable from users' web history and demographic features.

Harms. Harms in advertising include events such as non-consented outing [5], health privacy violations [6], advertising discrimination [7], and exclusion from economic opportunities [8]. Based on a collection of reports, and taking from the taxonomy of *Epistemic Fragmentation* [9], we define harm in the context of advertisement as: *(a) an inherently harmful ad (absolute harm); (b) a difference in ad exposure across user groups (contextual harm)*.

Existing literature focuses on verifying hypothesis, e.g., given a group of users and a harm, validate whether the harm exists. However, this approach might fail to detect context-dependent harms and prevents finding of harms that are not in the hypothesis. Instead, we follow an approach that *does not rely on predefined categories*. We rely on unsupervised clustering to create the groups of ads. We do not have restrictions: clusters can be based on semantic or visual patterns. We label each cluster with characteristics common to all ads in the cluster but not present in other clusters. We then measure exposure of groups of ads across groups of users. Limited by our dataset, we can only define user groups with the features of gender and age. We cannot study other groups, e.g., LGBT+ communities, because the information is unavailable. The mapping enables to detect harmful ads or a disparate delivery across user groups. Observing either pattern indicates a *potential* harm but context, i.e., information on ads and users they are served to, is needed to determine whether it is an actual harm. For example, weight-loss ads are harmful to people with eating disorders but otherwise harmless. Similarly, a manipulative ad is harmful regardless of the user but an ad for shoes is typically not.

3. Initial Results

From the 5 websites, we collect 704 personalized ads to the profiles we created. In parallel, we collect 699 *control* ads from the same websites, collected using a fresh browser without browsing history, thus without personalization.

Utility: Are the ads relevant? In Fig. 2, we plot on the left the distribution of likelihood that users click on a subset of control ads, e.g., not personalized; and on the right on ads personalized ads and served to them. We see an increase in likelihood. Notably, the "Likely" category sees a marked increase in frequency, overtaking "Unlikely". More precisely, we quantify this increase using the earth mover distance (emd) [10]. To establish a baseline, we evaluate utility of a non-overlapping subset of control ads. The emd shows that the likelihood of clicks is largely indistinguishable, with a score of 0.0308. In contrast, comparing the likelihood of click on control ads with the likelihood of click on personalized ads shows a larger distance with emd score of 0.1530, approximately five times the baseline noise. This approximates that users value personalized ads.

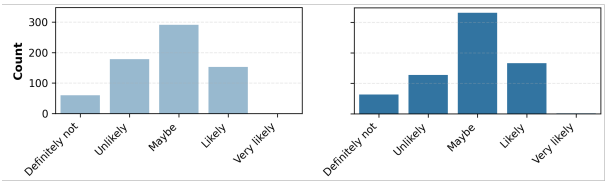

Figure 2: Click likelihood. Left: control; right: personalized

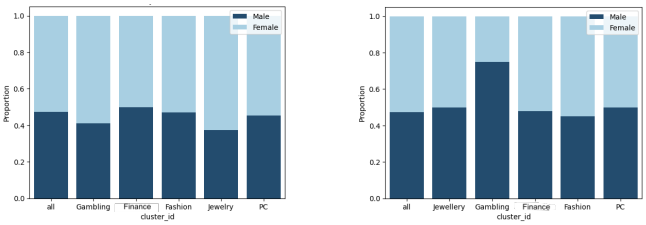

Figure 3: Gender distribution across advertisement clusters. Left: control; right: personalized.

Harmfulness: are the ads harmful? We cluster the personalized ads using Spherical K-means [11] and obtain 10 clusters. The clusters yield a VRC score of 153 and a silhouette score of 0.33, meaning that the clusters capture structure in the ads. We manually inspect, discard clusters where we do not observe a pattern, and label the remaining clusters: *fashion*, *PC/tech* (portable computer ads), *gambling* (casino and betting ads), *jewelry*, and *finance* (finance company ads).

We analyze age patterns across clusters, but find no meaningful shifts. Gender analysis, however, revealed a skew in gambling advertisements. As shown in Figure 3, over 77% of gambling ads were served to self-reported men's profiles. A chi-square test reveals a score of 2.6 and p-value of 0.11. The observed difference only has weak statistical significance, due to the small number of users. We find that only 4 out of 16 users who saw gambling ads had previously visited gambling websites, suggesting that 75% of gambling ad exposure occurred through some inferred information rather than direct signals on gambling interest. Furthermore, we find that users who saw gambling ads also visited more adult content websites on average (3.62 vs. 1.00) and slightly more financial service sites (1.75 vs. 1.40). The predominant male targeting (70.6% vs. 21.1% female) combined with limited correlation to gambling website visits indicate that the signal for showing these ads is not linked only to an interest in gambling.

4. Discussion

While based on a small sample size, our results offer valuable insights and demonstrate the potential of our methodology to find harms in real-world systems and measure the utility of ad campaigns. These results need to be validated with large-scale experiments and broader web exploration. A next step is to use the framework to study whether privacy-oriented solutions can truly protect consumers from harms without significantly reducing the relevance of ads.

References

[1] J. Kulshrestha, M. Oliveira, O. Karacalik, D. Bonnay, and C. Wagner, "Web routineness and limits of predictability: Investigating demographic and behavioral differences using web tracking data," *Proceedings of the International AAAI Conference on Web and Social Media*, 2021. [Online]. Available: https://zenodo.org/record/4757574

[2] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 1, pp. 92–112, 2015.

[3] G. A. Help, "About quality score for search campaigns," https://support.google.com/google-ads/answer/6167118?sjid=11992501053644637956-EU.

[4] P. Salem, C. Olsen, P. Freire, Y. Ding, and P. Saxena, "Tinytroupe: Llm-powered multiagent persona simulation for imagination enhancement and business insights," https://github.com/microsoft/tinytroupe, 2024, gitHub repository.

[5] T. Stobierski, "Facebook ads outed me," https://www.intomore.com/culture/you/facebook-ads-outed-me/, 2018.

[6] E. Zeng, X. Wu, E. N. Ertmann, L. Huang, D. F. Johnson, A. T. Mehendale, B. T. Tang, K. Zhukoff, M. Adjei-Poku, L. Bauer, A. B. Friedman, and M. S. McCoy, "Measuring risks to users' health privacy posed by third-party web tracking and targeted advertising," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ser. CHI '25. New York, NY, USA: Association for Computing Machinery, 2025. [Online]. Available: https://doi.org/10.1145/3706598.3714318

[7] L. Sweeney, "Discrimination in online ad delivery," *Communications of the ACM*, vol. 56, no. 5, pp. 44–54, 2013.

[8] European Digital Rights (EDRi), "How online ads discriminate," European Digital Rights, Tech. Rep., 2024, available at: https://edri.org/our-work/how-online-ads-discriminate/.

[9] S. Milano, B. Mittelstadt, S. Wachter, and C. Russell, "Epistemic fragmentation poses a threat to the governance of online targeting," *Nature Machine Intelligence*, vol. 3, no. 11, pp. 974–983, 2021.

[10] M. Mathey-Prevot and A. Valette, "Wasserstein distance and metric trees," 2021. [Online]. Available: https://arxiv.org/abs/2110.02115

[11] K. Hornik, I. Feinerer, M. Kober, and C. Buchta, "Spherical k-means clustering," *Journal of Statistical Software*, vol. 50, pp. 1–22, 09 2012.