PANOPTYKON FOUNDATION

Warsaw, May 31, 2023

Researchers' access to platforms' data – lessons learned from Panoptykon's case study "Algorithms of trauma"

(i) Description of our case study and the research questions

In 2021 we investigated the Facebook newsfeed of a young woman, mother of a toddler, who complained she had been exposed to disturbing, health-related sponsored content on the platform. Examples included in particular sponsored content (ads) with emphasis on cancer, genetic disorders or other serious conditions (such as crowdfunding campaigns for children or young adults suffering from these diseases). Since health had been a sensitive subject for her, this disturbing content fuelled her anxiety and was an unwelcome reminder of the trauma she had experienced in the past. At the same time, the woman claimed she had never explicitly and consciously provided any information pertaining to her (or her family's) health condition online or explicitly expressed her interest in that subject (not to mention consenting to use this information by any online platform, especially for advertising purposes).

We analysed the sponsored content in the woman's newsfeed and used this case study to better understand individual-level distribution of potentially "harmful" ads on Facebook which may have exploited the users' mental vulnerabilities. In particular we wanted to know:

- what was the proportion of health-related ads in the woman's newsfeed? Did those "harmful" ads constitute a significant/disproportionate fraction of the user's "ad diet"?;
- what was the role of ad targeting phase (i.e. where the advertiser specifies their target audience, by demographics, location, interests) and the role of ad delivery phase (i.e. the platform's optimisation of ads targeting - how the platform decides who among the targeted audience will actually see the ad) in the distribution of "harmful" ads?
- if an alteration of the ads settings offered by Facebook would have any effect on the user's experience (eg. would she get less health-related ads in her newsfeed?).

(ii) The data we have used

The woman donated her individual advertising and targeting data collected using a browser plugin. The data included the content of all the ads shown to her on Facebook, as well as the information revealed by the platform through the "Why Am I Seeing This?" transparency tool, i.e. a subset of targeting criteria selected by the advertiser of each ad. The browser plugin was a modified version of the Ad Observer software developed by researchers at the Center for Cybersecurity at New York University. Panoptykon supported by Piotr Sapieżyński, a research scientist at Northeastern University, analysed over 2000 ads in the woman's newsfeed over the period of 2 months. It turned out that approximately one in five of the ads presented to her were related to health, including a significant portion of such ads featuring terminally ill children or references to fertility problems. We have also found 21 health-related targetable "interests" which Facebook assigned to her in order to personalise ads, including: "oncology", "cancer awareness", "genetic disorder", "neoplasm" and "spinal muscular atrophy". All of these interests were inferred by the platform based on the user's online activity on and off Facebook, and subsequently made available to advertisers to target their ads.

As part of the experiment, the woman tested all available Facebook settings to change her behavioural profile (in particular – she disabled the health-related "interests"). It turned out that, although Facebook made some ad control tools available, none of them were effective in decreasing the prevalence of the most problematic ads in the woman's newsfeed. Unfortunately, the user's experience hardly improved when she changed her settings. The number of "harmful" ads was changing during the experiment but after 2 months it returned to nearly the original level.

Disabling health-related interests did remove the ads targeted using these interests by the advertisers. However, two other phenomena occurred, which effectively cancelled out any perceived change introduced by disabling these interests. *First*, Facebook inferred "new" interest categories (such as "Intensive Care Unit", "Preventative Healthcare", and "Magnetic Resonance Imaging") and assigned them to our user for targeting by the advertisers. The content of the ads targeted using these criteria was not substantially different from the ads targeted through the disabled interests and hence the user's experience did not improve. *Second,* the user started receiving health-related ads that were *not targeted* using any sensitive interests by the advertisers. Why would the user see these ads despite the lack of targeting? Facebook and other online advertising platforms *optimize the delivery of ads* to those users among the targeted audience who are predicted most likely to find the ads relevant based on the *content of the ad*, rather than just targeting parameters. Facebook had previously inferred that our user was likely to engage with health-related content, and thus showed her such ads even when the advertisers did not explicitly requested it.

(iii) Results of the research and what we *could not* verify based on the data we had

We obtained evidence that:

I. Specific type of sponsored content (health-related ads) fuelling the woman's anxiety were disproportionately prevalent in her newsfeed.

This may suggest that algorithm-driven distribution of ads may expose certain vulnerable users to large amount of sponsored content that has harmful consequences for them and that the platform's ads distribution mechanisms exploit inferred traits, sometimes of a highly sensitive nature, which users have not willingly disclosed (*hypothesis 1*) .

Moreover results of the case study suggest that:

II. Facebook's optimization algorithms played a crucial role in delivering sponsored content that may have led to serious individual harm (*hypothesis 2*).

III. Affected individual had no effective tool to control this process. Disabling sensitive interests in ad settings limits targeting options for advertisers, but does not affect Facebook's own profiling and ad delivery practices (*hypothesis 3*).

Even though our experiment provides strong indications for hypotheses 1-3, we were unable to ultimately verify them due to limited scale of this case study (analysing single person's newsfeed) and, more importantly, lack of access to algorithms and data required for, in particular, a thorough study of the implications of platform's controlled phase of ads distribution process, i.e. ad delivery.

Moreover, neither the current (nor foreseen in the DSA) ad transparency tools (to be) provided by platforms are adequate for an in-depth studying of ad delivery as they are focused on the ad targeting phase.

(iv) What data we think are needed (should be shared with regulators and researchers) to verify our hypotheses

To enable the scrutiny of platforms' personalised ads distribution systems and their real-life implications, researchers need ***more information about the operation of ad delivery algorithms***, including those involved in: running the auction, relevance measurement and estimation, and bid and budget allocation on advertisers' behalf.

Such data access should allow to:

- disentangle the roles of targeting and delivery optimization in shaping the exposure of individuals to ads;
- investigate why some users are exposed to more harmful experiences than others;
- measure demographic biases in delivery of opportunity advertising (housing, employment, education, etc.);
- understand factors that play a role in ad delivery optimization algorithms and lead to skew in delivery, including factors related to content of the ad, user data (eg. geographical, demographic), user reactions (ways users can engage with ads, eg. likes commenting, sharing) or telemetric information (eg. how long each user spent looking at the ad);
- assess effectiveness of the user ad control settings provided by the platform.

In particular, access to following data is needed:

- all targeting criteria chosen by the advertiser for each ad;
- content classification derived from the ad (for example ad topic);
- summary demographic statistics of the audience eligible to see the ad based on the targeting criteria, including those arising from the use of Custom Audiences, Lookalike Audiences, and similar audience products;
- summary demographic statistics of the actual audience who were shown the ad;
- summary description of behavioral factors by which delivery was optimized (for example "preferentially shown to users who recently interacted with health-related content").

The privacy concerns that arise from publishing this information are all related to the fact that platforms allow for public interactions with sponsored content: when user A "likes" an ad targeted by parameter X, the inferred association between A and X becomes public. Platforms should remove the ability to interact with targeted content to ensure that the proposed transparency measures do not come at the cost of loss of privacy of individual users.

***

For any further questions regarding this submission, please contact our team members:

Katarzyna Szymielewicz (katarzyna.szymielewicz@panoptykon.org)
Dorota Głowacka (dorota.glowacka@panoptykon.org)