PANOPTYKON FOUNDATION

February 2026

Breaking the Loop: Exploring User Agency in Escaping Algorithmic Rabbit Holes on Social Media

The early-stage insights from Panoptykon's and Piotr Sapieżyński's (Northeastern University) study

Background

In the summer of 2023, we analysed the Facebook feed of a user who had reported a high volume of distressing posts appearing in the "Suggested for you" section – content related to health issues, tragic accidents, and death. The resulting case study, titled Algorithms of Trauma 2 confirmed a significant prevalence of such material in the user's feed: over a period of nearly two months, 56% of suggested posts contained themes the user identified as problematic, amounting to an average of approximately 27 such posts per day. Furthermore, the study found that the platform's explicit feedback tool – namely the "Hide post – See fewer posts like this" button – was ineffective and failed to produce the expected mitigating effect. Despite the user clicking the button 122 times during the study period, the frequency of problematic "Suggested for you" content did not decrease. In fact, not only did the intervention fail to reduce exposure to unwanted posts, but the frequency of such content slightly increased following the user's actions (see details in: D. Głowacka, K. Szymielewicz, P. Sapieżyński, Algorithms of Trauma #2. Stuck in a "doomscrolling trap" on Facebook? The platform will not let you escape, 2023).

Our case study aligned with other reports suggesting that engagement-driven algorithms on social media platforms may create self-reinforcing feedback loops, gradually steering users toward increasingly narrow and potentially disturbing content. This dynamic can result in significant individual harm, including adverse effects on users' mental and emotional well-being. Moreover, our findings indicate that individuals affected by these mechanisms often lack effective tools to meaningfully control the content recommended to them, making it difficult to escape toxic algorithmic "rabbit holes."

In 2025, together with Piotr Sapieżyński from Northeastern University, we initiated a new research project: *Breaking the Loop: Exploring User Agency in Escaping Algorithmic Rabbit Holes on Social Media*. It has been focused on evaluating the effectiveness of explicit feedback tools currently available on Facebook. Building on our previous methodology and data collection tools, we expanded the scope of the study to include a slightly larger-scale analysis.

Through our findings, we aim to shed further light on whether, now that the DSA is fully operational, the explicit feedback tools offered by VLOPs genuinely deliver on their promises – and to place this analysis within the broader discourse on the enforceability of the DSA, including potential avenues for individual redress, as well as the enforcement of measures aimed at mitigating systemic risks.

Existing research

Existing independent research on platforms' feed control tools, following the entry into force of the DSA, remains limited, particularly in Europe. The available studies present mixed findings but often highlight persistent shortcomings in the effectiveness, availability, and granularity of these mechanisms (see excerpts in the table below). At the same time, a notable share of this research relies primarily on user interviews or self-reported experiences rather than systematic monitoring of actual content feeds, which further weakens the evidence base. Consequently, despite increasing regulatory attention, our empirical understanding of how these tools function in real-world conditions remains incomplete.

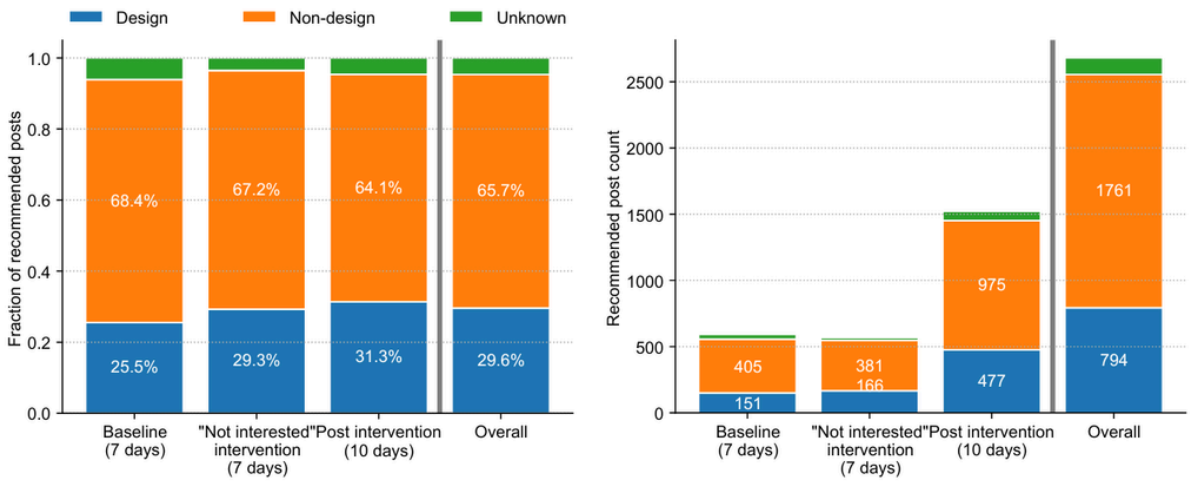
TikTok	Most participants said they **did not think their "Not Interested" was effective** or felt that its effectiveness was inconsistent. Users complain feedback is **not granular enough**. (...) Users often **struggle to express their preferences precisely** through such simple interactions. These features are typically designed as one-tap buttons with which users struggle to describe the reasons for their feedback. (...) After submitting the feedback and observing the changes in the personalized feed, [users] were often uncertain of how each of their interactions was reflected in the algorithm. **(J. Hong, E. Ko, J. Kim, J. Jang, 2025)** Once captured in a content bubble, users found it difficult to leave, spending long periods of time being recommended undesired content **("algorithmic persistence")** (...) While the filtering function technically worked such that the set of specified keywords stopped appearing on their FYPs, **they still saw content from the same genre, all with hashtag content that had not been blocked.** **(J. Vera, S. Ghosh, 2025)**

Facebook	We find that utilizing the **"See less" ad control for the topics we study does not significantly reduce the number of ads shown by Meta on these topics**, and that the control is less effective for some users whose demographics are correlated with the topic. **(J. Castleman, A. Korolova, 2024)**
YouTube	**Using the "Not interested" button worked best,** significantly reducing [recommended videos dedicated to the assigned topic on the user's homepage] in all topics tested, on average removing 88% of them. **Neither the stain phase nor the scrub phase, however, had much effect on videopage recommendations** (those given to users while they watch a video). **(A.Liu, S. Wu, P. Resnick, 2023, updated 2024)**

At the same time, we are aware that researchers are increasingly scrutinizing feed-control tools, including a forthcoming project by Piotr Sapieżyński et al. that focuses on TikTok. In the paper *When 'For You' Isn't For You: Measuring User Agency in TikTok's Algorithmic Feed* (work in progress, not yet published, to be presented at ICWSM 2026), L. Kaplan, D. Patel, N. Gerzon, A. Mislove and P. Sapieżyński observe that tapping "Not interested" button on TikTok does indeed reduce the amount of content on a given topic recommended to a user. However, once the user stops selecting "Not interested," the platform gradually resumes recommending the previously unwanted content. If the user reverts to watching these videos again, the feed quickly repopulates with them to the same extent as before the intervention. The authors refer to this phenomenon as the *relapse effect*. These findings are based on sock-puppet experiments and are yet to be replicated with real users.

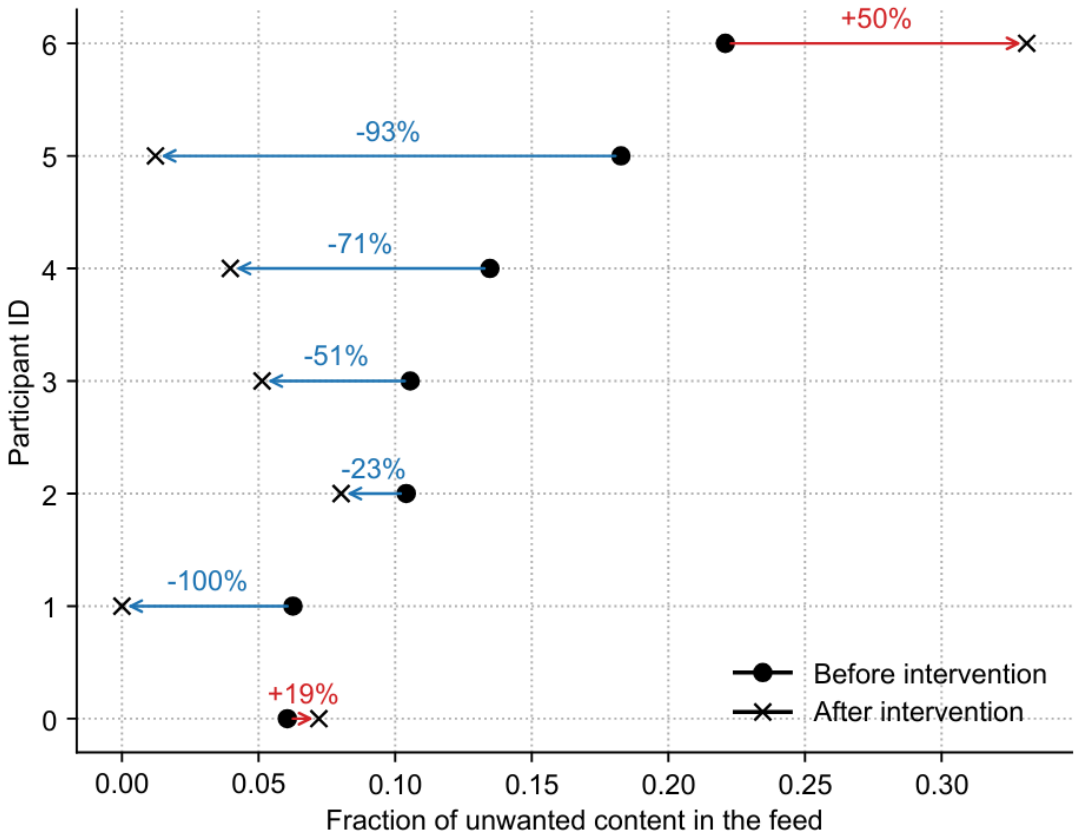
The early-stage insights from Panoptykon's study

Between 18 August and 10 September 2025, as part of our preliminary work, we conducted a pilot single-participant case study that closely replicated the methodology used in *Algorithms of Trauma 2*. In this iteration, we examined the Facebook feed of an individual who reported being exposed to a substantial number of "Suggested for you" posts related to design and architecture. The key finding of this pilot study is that clicking more than 100 times on the "Not interested. Less of your posts will be like this" button (between 25 and 31 August) did not produce any meaningful or lasting reduction in the prevalence of such unwanted content in the user's feed (see the graph below for details).



Between 18 and 27 December 2025 we conducted another Facebook-focused study on a larger sample of seven participants. In this iteration, we slightly adjusted the methodology: instead of asking participants to identify unwanted content upfront, we first observed their feeds during a passive-scrolling phase (5 - 9 December) and used automated tools to map their individually emerging "rabbit holes" by clustering all "Suggested for you" posts (10-12 December). Only after presenting these clusters to participants did they select the topics they wished to eliminate, which they then marked as "Not interested" in the intervention phase (13-17 December). In the final phase (18–27 December), we monitored participants' feeds following the intervention.

The results differed from our earlier pilot: on average, clicking "Not interested" reduced the share of targeted content by 38%, with statistically significant effects for five of the seven participants. For two individuals, however, the amount of unwanted content either increased or did not meaningfully decrease (see details in the graph below).



On the Y-axis, each participant is represented individually, while the X-axis shows the proportion of "suggested for you" content related to the unwanted topic. Every participant is visualized with a separate arrow: blue when the share of unwanted content decreased after using "Not interested", and red when it increased. The final participant at the bottom shows an increase, though this change is not statistically significant. In the chart, "o" denotes the baseline measurement combined with the clustering phase (5–12 December), whereas "x" marks the end of the passive-scrolling phase (27 December).

The results may suggest that the effectiveness of the "Not interested" feature has improved. However, further analysis is needed to understand why its impact varied across participants and to assess the extent to which these effects are sustainable over time – that is, whether the reduction in unwanted content persists or whether such content eventually reappears in users' feeds. This question is particularly important given indications from our earlier case studies, as well as from the ongoing research with TikTok sock-puppet accounts mentioned above, which suggest that the relapse effect is plausible and may systematically undermine the long-term efficacy of user-driven feed control mechanisms. Interestingly, for more than half of the topics that were not marked as "Not interested", their presence in the feed also declined following the intervention. This may suggest that the platform's response to user feedback may operate at a broader level than the specific content category targeted by the user.

Therefore further data collection and analysis are needed to clarify in particular the following issues:

- the extent of the relapse effect,
- the extent to which observed changes in the feed are attributable to drift, i.e., temporary fluctuations in topic popularity rather than user-driven signals,
- the degree of alignment between the effects of the "Not interested" button and users' actual expectations, particularly given the limited precision of current controls (e.g., users have limited possibility to indicate that they wish to retain health-related content while excluding dieting-related content).

Possible next steps for this research agenda include:

- measuring the 'relapse effect': a follow-up to the pilot study to assess whether unwanted content reappears in users' feeds over the longer term (we are planning to revisit our participants' feeds as soon as we update our data collection tool that in the meantime was disabled by the platform),
- conducting a comprehensive, large-scale investigation: a systematic examination of content feeds across multiple (50 +) real-user Facebook accounts to validate and extend the initial insights obtained so far.

Relevance of the study for the DSA's enforcement

We believe that under the DSA users should have access to effective tools that allow them to meaningfully control the content recommended to them, enabling them to escape toxic algorithmic "rabbit holes" when necessary. Such tools are essential for strengthening user agency and reducing the risks associated with hyper-personalised content recommendations.

Our central argument is that, in principle, VLOPs deploying explicit feedback tools that appear insufficiently accessible, ineffective or misleading from the user's perspective may be in breach of several provisions of the DSA – specifically Article 27(3), in conjunction with Articles 25 and 28 (particularly in light of the European Commission's guidelines on the protection of minors), as well as the obligation to implement adequate mitigation measures under Article 35. The question of compliance with art. 35 is particularly relevant given that most social media VLOPs have themselves identified such tools as part of their risk mitigation strategies in their annual risk assessment reports, yet have failed to provide any data demonstrating that these mechanisms actually produce the intended mitigating effect.