

Algorithms of Trauma #2

AUTHORS

Dorota Głowacka &
Katarzyna Szymielewicz
(Panoptykon Foundation),
Piotr Sapieżyński
(Northeastern University)

CONTENT

1	Summary
1	Introduction
2	The participant
3	Study timeline
4	Content labelling
4	Summary of findings
6	Other studies questioning the effectiveness of VLOPs' human agency tools
7	Relevance of the study for the DSA's enforcement
11	Conclusions. Safety buttons: a low hanging fruit to improve user experience
11	Limitations
12	Annex. Examples of problematic "Suggested for you" content in the participant's feed
16	Bibliography

Stuck in a "doomscrolling trap" on Facebook? The platform will not let you escape

Summary

"Algorithms of trauma #2", a new study from Panoptykon Foundation, tested Facebook's recommender system and the effectiveness of one of the human agency tools provided by the platform to influence user feed. The study analysed the feed of the user complaining about large amount of distressing posts displayed in the "Suggested for you" category, pertaining to health problems, tragic accidents, and deaths. The study confirmed a significant prevalence of such content in the user's feed: over almost two months 56% of suggested content (1416 posts) included themes that the user labelled as problematic. This amounts to an average of around 27 such posts per day.

Moreover the study found that the explicit feedback tool provided by the platform, i.e. the "Hide post – See fewer posts like this" button, was ineffective and did not have the expected mitigating impact. Clicking on the button 122 times in the course of the study did not lead to lowering the frequency of the problematic "Suggested for you" content. In fact, not only did it not free the user of unwanted posts, but the frequency of such posts slightly increased after the user's intervention.

We argue that, under the Digital Services Act, VLOPs deploying such "deceptive" explicit feedback tools which, from a user perspective, do not have any positive impact on their feed, may violate obligations imposed on online platforms under art. 27.3 in conj. with art. 25, as well as their obligation to implement adequate mitigation measures stipulated in art. 35 of this Act.

Introduction

Whenever a user scrolls through their Facebook's feed, they are presented with a selection of content that has been algorithmically curated based on the traits and interests attributed to them by the platform with the stated goal of maximizing user satisfaction. The sources of presented content can be divided into three categories:

- ▶ organic sources that the user explicitly subscribed to (their friends, groups, and Pages they followed),

ACKNOWLEDGEMENTS

Authors would like to thank Joran van Apeldoorn (Bits of Freedom) for his expert contribution to this report.

- ▶ non-organic (sponsored) sources (from advertisers who included the user in their target audience),
- ▶ organic sources that user did not subscribe to (so called “Suggested for you” content); users cannot independently pre-define the subject matter in which they are interested to determine the contents of suggested posts.

Users can choose to unfriend or mute the organic sources they subscribed to if they wish not to see their content. However, influencing the selection of ads might prove more difficult. Our [previous study](#) (Panoptykon & Sapieżyński, 2021) showed that the ad controls fail to influence the selection of presented ads in a meaningful way – despite the user disabling all “topics” and “interests” related to parenting and health, the portion of their ad diet that pertained to these topics remained unchanged.

This study focuses on “Suggested for you” category. It aims to shed more light on the functioning and potential negative effects of Facebook’s algorithms recommending “Suggested for you” posts and to measure how much agency users have over this category of content. The platform offers tools for the user to indicate whenever they are displeased with the particular post they are suggested. But to what extent do they actually change the prevalence of unwanted content in their feeds?

To find out, we monitored the Facebook’s feed of one of the users who complained to Panoptykon about large amounts of distressing posts displayed to her in the “Suggested for you” category, which referenced health problems, tragic accidents, and deaths (see screenshots in the Annex attached below). After monitoring the participant’s feed while she routinely used Facebook for a certain period of time, we had the user flag all health, accident, and death-related content that was suggested to her as “unwanted”, and then measured its prevalence in the weeks following the intervention.

The participant

Our participant is person in her 30s and a mother of a toddler. Before she had her baby, one of her loved ones had suddenly fallen ill and died of cancer. It was then that she began to suffer from intense anxiety over the life and health of those closest to her, and her own. The problem has only exacerbated after the baby’s birth (as she started being concerned about baby’s health as well). The participant complained about living under constant stress and experiencing several psychosomatic symptoms. She struggled with compulsory “googling symptoms” and consuming large amounts of health-related content online, mainly off-platforms (a habit she was trying to change when realized it being one of the drivers behind escalating her anxiety¹). As regards the user’s activity on Facebook, she has been using it mainly in a passive manner, i.e. to access information; if she ever shared any content or a reaction on the platform, it was never related to her physical or mental health. She did join however several support groups for sufferers of chronic anxiety on the platform. In 2021 she also started psychiatric and psychological treatment.

The problematic “Suggested for you” posts started plaguing her feed ca. a year ago. According to the participant, an exposition to this content caused a lot of distress and fear, and has become a factor contributing to once again exaggerating her health anxiety, which made her want to cut herself off from this type of posts.

1. There is also research confirming the that “googling symptoms” and exposition to the health-related content online increase risk for developing or exacerbating health anxiety, as well as lead to unnecessary costs in time, distraction, and engagements with medical professionals. See for example (White & Horvitz, 2009).

INTERVIEW WITH THE PARTICIPANT ►

To hear the full story, listen to the special episode of the [“Panoptikon 4.0” podcast](#) dedicated to the “Algorithms of trauma #2” study



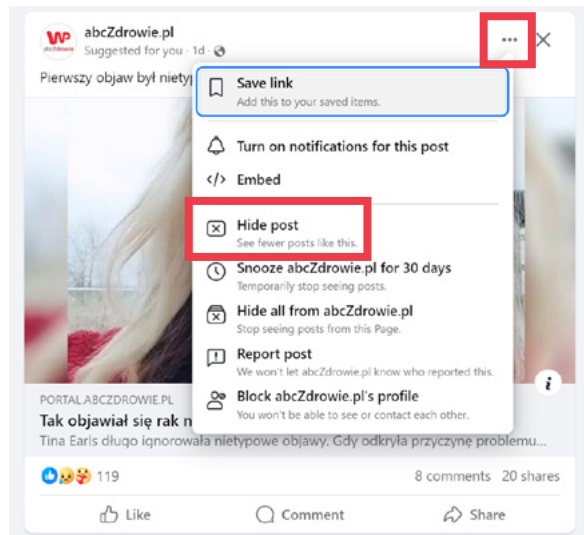
“(…)[W]hen I see a post on Facebook where someone suffers from a disease, I find it hard not to take it personally. I just relate to it immediately. I start to scan my body, my brain shifts into high gear, I analyse if I’m not having any of the symptoms. The same things happen with stories about children’s health: I start analysing my child right away. And, more often than not, I will find something to worry about. My anxiety then goes through the roof, and I get the temptation to learn more about the disease, so I start reading, googling, researching online. This, in turn, probably causes Facebook to plant even more of this information in my feed. As a result, my worries escalate, the visions of disease seem more and more realistic, I get the feeling that I should have it tested, I start looking up doctors… And I’m on a downward spiral of fear.”

Study timeline

The data collection lasted between 7th of June and 28th of July 2023. The study was divided into three stages:

1. **Baseline observation**, 7th of June till 22nd of June, with two days of data missing. This stage was to establish a baseline frequency at which the participant had been shown problematic content through the “Suggested for you” posts on Facebook. Data collection and interacting with the platform was facilitated through an automated browser which scrolled through the feed, as well as saved, and categorized each post as organic/sponsored/suggested. In this period the participant would collect 100 top posts presented in her feed a few times a day (0-6, on average 2.6 sessions a day). Each post was presented on the screen for the same short period of 1 second.
2. **Active intervention**, 23th of June and 2nd of July, with two days of data missing. This stage was for the participant to indicate to Facebook what content she finds problematic and would wish to see less of. The scraper would pause for 20 seconds every time a piece of “suggested” content appeared on the screen and the participant could choose to “Hide post - See fewer posts like this” by tapping a button in the top right of each of posts (see the screenshot below). Whenever she did, Facebook would hide the post and the scraper would refresh the page and continue scraping until a total of 100 pieces of content were displayed. **The participant tapped the button 122 times to flag unique pieces of undesired content.**

“HIDE POST - SEE FEWER POSTS LIKE THIS” BUTTON ON FACEBOOK* ►



*Depending on the device and the version of the app, the explicit feedback tools available on the platform may slightly differ.

3. **Evaluation of intervention effects**, 3rd of July until 28th of July, with four days of data missing. This stage was to verify whether the intervention led to less problematic content being suggested to the user. The scraper would no longer pause the suggested content.

Content labelling

Within the study period (ca. 2 months) the participant was “suggested” over 2500 unique pieces of content, corresponding to approximately 22% of all displayed stories. She subsequently labelled each suggested story as “problematic” or “not-problematic”, based on whether the content was referencing health problems, tragic accidents, and death. 15%, or 373, pieces of content were additionally labelled by a different annotator with a consensus score of 97.8%.

Summary of findings

The problematic “Suggested for you” content had a significant prevalence in the participant’s feed. Over 56% (1,416 posts) of the suggested posts displayed to the participant were labelled as problematic (there were days however when even 8 out of 10 suggested posts would fall under this category). Ultimately, almost every eighth post in her entire newsfeed turned out to be problematic suggested content. That is an average of ca. 27 problematic suggested posts per day.

The participant’s intervention, i.e. clicking 122 times on the “Hide post – See fewer posts like this” button in the course of the study, in the long perspective did not lead to lowering the frequency of suggested content in general, nor the frequency of the problematic content in particular. In fact the frequency of problematic “Suggested for you” content slightly increased after the intervention.

Over the week that the participant tapped the “Hide post – See fewer posts like this” button, the overall number of suggested posts did decrease marginally. However, at the same time, the ratio of problematic content to all suggested posts increased (i.e. there was a greater concentration of problematic posts). Within a few days the total number of suggested posts returned to its former level, and then continued to grow, eventually exceeding the level from the beginning of the study.

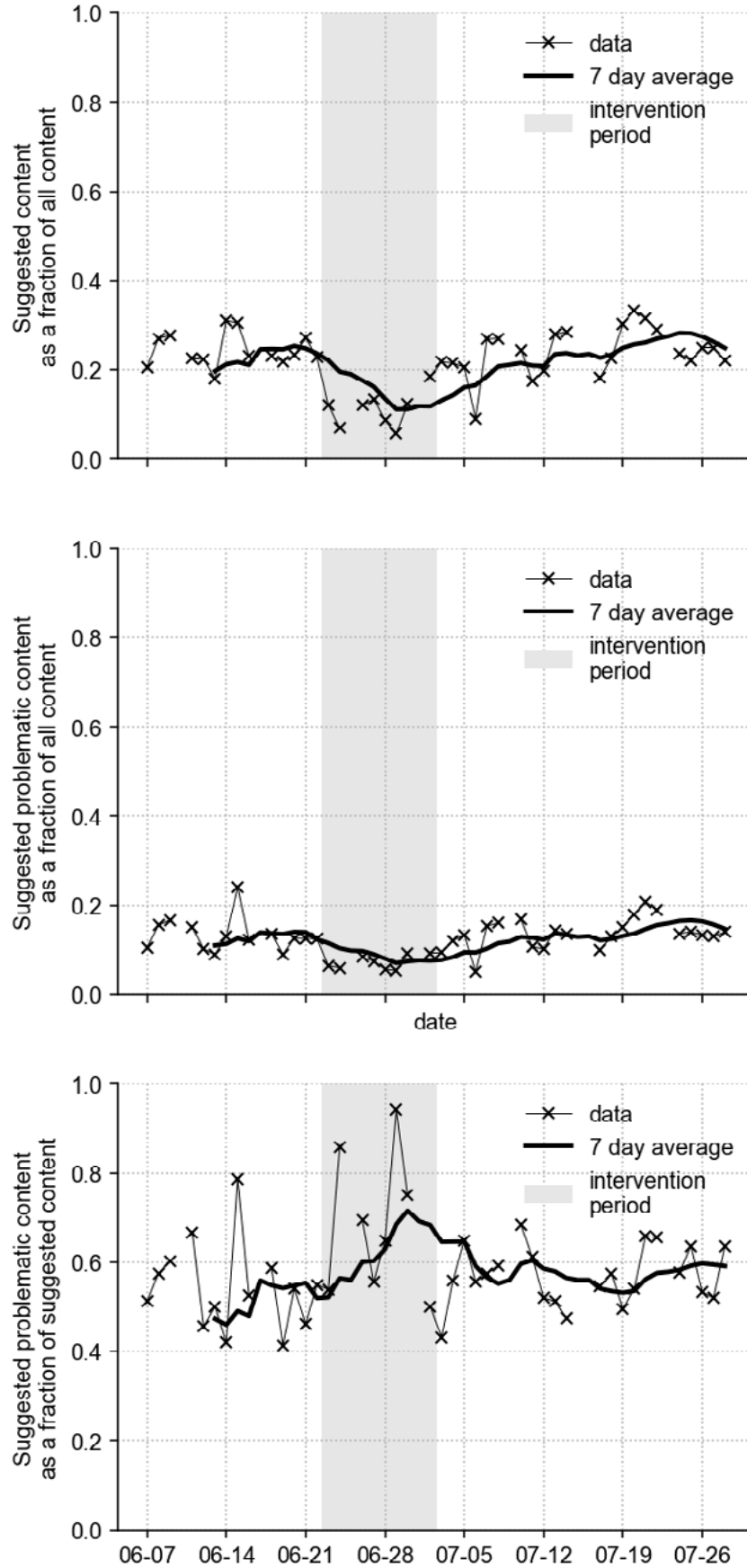
TABLE 1 ►

Prevalence of “Suggested for you” content and problematic “Suggested for you” content in the participant’s feed

	Before intervention	After intervention
Suggested content as fraction of all content	24.1%	24.3%
Problematic suggested content as fraction of all content	12.9%	13.8%
Problematic suggested content as fraction of suggested content	53.7%	56.6%

DETAILED TIMELINES ▶

Each plot corresponds to a subsequent row in Table 1



Other studies questioning the effectiveness of VLOPs' human agency tools

Panoptykon's study is not the only one to question the functioning of the control tools provided by very large online platforms (VLOPs) to their users with the promise that they can "impact their feed". Several papers and reports have shown that explicit feedback and expressed preference for specific sources or topics had little or no impact on recommended content.

FACEBOOK

The already mentioned previous study by the Panoptykon Foundation (Panoptykon & Sapieżyński, 2021) indicated that changes in advertising preferences did not allow the user to permanently limit the ads related to the "interests" attributed to them by the algorithm. This study concerned the same participant who at the time complained she had been exposed to disturbing, health-related sponsored content on Facebook. Examples included in particular ads with emphasis on cancer, genetic disorders or other serious conditions (such as crowdfunding campaigns for children or young adults suffering from these diseases). Panoptykon analysed over 2000 ads in the participant's feed over the period of 2 months between mid-May and mid-July 2021. It turned out that approximately one in five of the ads presented to the participant was related to health. As part of the experiment, the participant tested all available Facebook settings to change her behavioural profile. In particular – she disabled over 20 health-related "interests" that Facebook attributed to her for the purpose of personalising ads, including for example "Cancer", "Genetic disorder" or "Hospital". It turned out that, although Facebook made some ad control tools available, none of them were effective in influencing how algorithms select content and maximise user engagement. The participant's experience hardly improved after she had changed her settings. The number of disturbing 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 the removed interests, but – on the one hand – new categories were inferred, such as "Intensive Care Unit", "Preventative Healthcare", and "Magnetic Resonance Imaging", and – on the other hand – the topic of problematic ads and their prevalence remained unchanged even when the sensitive "interests" were no longer used. The results suggested therefore that it was the platform's ad delivery algorithms, and not the targeting criteria selected by ads sponsors, which played a crucial role in distribution of sponsored content, while at the same time ad delivery algorithms oriented for user engagement may lead to serious individual harm.

YOUTUBE

According to the study published by the Mozilla Foundation (Mozilla Foundation, 2022) also the YouTube users do not have significant influence over the recommendations provided to them. For many users using the content feedback tools available on the platform (such as the "Not Interested", "Dislike" or "Don't recommend channel" buttons) was simply frustrating. To examine how YouTube's recommendation algorithm handles user feedback, the authors combined qualitative and quantitative research methods. First, they surveyed 2,757 participants to better understand their experiences with YouTube's recommendation algorithm. They found that a significant minority (39.3%) of the respondents who used YouTube's control tools did not feel that doing so impacted their recommendations at all. Just over a quarter (27.6%) felt that their recommendations did change in response, and fewer (23.0%) felt the system had an ambivalent or mixed response (for example, within that group 23.7% users noticed some positive changes after they used the control tools, but said that over time unwanted recommendations would return). Second, the authors of the study analysed data donated by 22,722 people, generating a dataset of the 567,880,195 videos they were recommended. The quantitative research largely validated the results of qualitative analysis, confirming that the feedback tools provided were inadequate for preventing unwanted recommendations. Even the most effective user controls ("Don't recommend channel") prevented less than half of "bad

recommendations” (43%), while tools such as “Not interested” or “Dislike” had even less impact (prevented only 11% and 12% of “bad recommendations” respectively).

A more recent study focusing on YouTube’s control tools (but using a different method – a sock puppet audit) provided more mixed results, concluding that different strategies to remove unwanted content on YouTube work to different degrees (Liu et al., 2023). In this study, the “Not interested” button was found to be the most effective in reducing unwanted recommendations from the platform’s homepage, on average removing 88% of them. At the same time none of the control tools tested had much impact on the platform’s videopage recommendations (those given to users while they watch a video).

TIKTOK

A study analysing the user perception of content feedback tools has been also conducted among TikTok users (Milton et al., 2023). “Many participants” claimed that the control tools on the platform did not allow them to effectively tailor the content of the TikTok’s “For You Page” (FYP) to their preferences. According to the study: “several participants were frustrated by the fact that <<Not interested>> [button] didn’t seem to work as they thought (...). The combination of lack of control over one’s feed and the continued delivery of unwanted content creates a problematic situation where the TikTok’s FYP is perceived to disregard a user’s preferences”. The authors further stressed that: “in the most extreme cases, participants felt like the lack of control of the FYP led to harmful consequences to their well-being. We describe this as the FYP acting like a runaway train, a technological system that users cannot control but feel that they cannot leave or disengage from.”.

Relevance of the study for the DSA’s enforcement

Until recently users frustrated with ineffective explicit feedback tools on social media platforms in reality had no avenues for legal redress and could only rely on the platform’s willingness to follow good practices and “fix” their dysfunctional features. The situation has changed after the EU adopted binding rules for social media platforms.

25th of August 2023 marks a date by which all the VLOPs were expected to comply with their obligations according to the Digital Services Act (DSA), as well as identify systemic risks and implement adequate mitigation measures according to art. 34-35 of this Act. Even though our study had been carried out shortly before the new regulation became fully effective with regard to VLOPs, we believe it indicates areas, which Meta (and possibly other VLOPs) should address to ensure compliance with the new regulatory requirements. We hope that the European Commission will verify whether problems illustrated by our case study have been solved.

In terms of compliance with the DSA, the following issues call for investigation:

1. ineffectiveness of human agency tools provided by VLOPs which do not enable the actual control over the recommended content (as it may constitute a violation of art. 27.3 in conjunction with art. 25 of the DSA);
2. significant ratio of recommended content triggering mental health issues in the user feed, combined with ineffectiveness of human agency tools when it comes to limiting users’ exposition to content they find disturbing (as it may infringe upon VLOPs’ obligation to implement adequate mitigation measures according to art. 35 of the DSA).

RULES ON RECOMMENDER SYSTEMS AND A PROHIBITION OF “DARK PATTERNS”

Ad. 1. DSA has introduced new requirements for recommendation algorithms used by all online platforms. In principle, DSA does not require platforms to provide users with human agency tools to modify their default feed (incl. buttons such as “Hide post – See fewer posts like this”). But once the platform decided to provide such tools, they should be clearly described in the terms and conditions, easily accessible and, most importantly working effectively at any time. These obligations stem from article 27.3, supported by the prohibition of so called “dark patterns” (manipulative design practices to influence user behavior) in article 25 of the DSA.

By offering tools for explicit feedback, Facebook creates an impression that users can influence content that will be recommended to them, at least by eliminating what they find disturbing. Providing the “Hide post – See fewer posts like this” button, as well as its description in the Facebook’s Help Center², suggest that taking the effort to flag the undesired content should lead to limiting the user’s exposition to it. Our case study shows that it was clearly an empty promise: the platform ignored explicit feedback from the user and continued to recommend content flagged as unwanted.

SYSTEMIC RISK ASSESSMENT AND MITIGATION MEASURES

Ad. 2. DSA obliges VLOPs to periodically identify, analyse and assess the systemic risks stemming from the design or functioning of their services, including “the design of their recommender systems and any other relevant algorithmic system” (art. 34.2a). Among other factors, risk assessments should take into account “actual and foreseeable negative effects of the functioning of recommender systems for the exercise of fundamental rights” (including right to privacy) (art. 34.1b), protection of public health and serious negative consequences to the person’s physical and mental well-being (art. 34.1.d). On the basis of their risks assessments, VLOPs should adopt reasonable, proportionate and effective mitigation measures such as “testing and adapting their algorithmic systems, including their recommender systems” (art. 35.1e), as well as “adapting the design, features or functioning of their services, including their online interface” (art. 35.1a). VLOPs should have completed their first risks assessments and implementation of corresponding mitigation measures by 25th of August 2023.

We argue that when providing ineffective feedback tools, such as “Hide post – See fewer posts like this” button, VLOPs violate their obligations under article 35 of the DSA. Ineffectiveness of such tools has particularly detrimental consequences for users who are “haunted” (nagged) by content which they perceive as harmful. In such cases VLOPs should not only facilitate agency and respect choices of their users when shaping the feed but, above all, they should provide them with effective “safety buttons” to mitigate experienced harms. In the context described in our case study, an effective feature removing unwanted recommendations from the feed should therefore not be an option, but a mandatory tool VLOPs should make available to their users as a necessary mitigation measure.

2. In the Facebook’s Help Center, in the section “Learn about and manage suggested content in your Facebook Feed”/“Manage what you see”, Meta assures its users that they “have control over which content is suggested for them”. Among other options it provides “to manage what users see”, they can select “Hide post to see less content like this”. Moreover the company explains in its policy that, in principle, its goal is “to avoid making recommendations that could be low-quality, objectionable, or particularly sensitive” or “recommendations that may be inappropriate for younger users” (“Facebook’s standards for suggested content in Feed”) (Meta, 2023).

Even though our study only investigated the feed of one person and further research is needed to confirm that the observed lack of user agency represents a more general problem, we argue it should be considered a “systemic risk”. First of all, this is because ineffectiveness of user feedback tools has been observed by researchers on other social media platforms (beyond Facebook; see the previous section). Second, our study resonates with the pre-existing research suggesting that over-exposition of vulnerable users to (what users themselves perceived as) triggering content is “not a bug but a feature”, ingrained in the business model of dominant social media platforms.

It has been established that the fact that recommender systems are optimised to maximise user engagement is linked to feedback loops that drive users into narrower selections of content, including so called toxic “rabbit holes” or

“doomscrolling traps”, which are difficult to escape (Panoptykon et al., 2023 [1]). Research has shown that composition of the feed intended to engage users at any cost may cause real damage to their mental health: apart from exacerbating anxiety, such experience may also fuel suicidal thoughts or disordered eating (Ibidem). Content triggering anxiety in some users often will not be dangerous as such, and therefore, will not be eligible for moderation. While it can be acceptable and appropriate in isolation, it becomes harmful if consumed too much or by vulnerable users. Therefore the most adequate mitigation measure is to offer “safety buttons” or “brakes” designed for vulnerable users if and when they experience the pattern of scrolling for negative information.

From explanations of how recommender systems work published by VLOPs (as part of their compliance with their transparency obligations under the DSA) as well as independent experts (e.g. Integrity Institute, 2023) we also know that the overarching objective to maximise user engagement implies other design choices, including default notifications and what signals are taken into account by the ranking algorithm. Social media platforms, as we know them, attribute more weight to behavioural observations and downplay explicitly expressed user preferences³. In this context, it is not surprising that users feel disappointed when they try to use control tools and discouraged from customizing their experience (Smith et al., 2021). The end result is what we see in data (eg. Jin et al., 2017): great majority of users falls back on the default settings, reinforcing VLOPs’ current business model and existing default feeds’ structures.

That is why the problem highlighted in our case study can’t be solved without systemic solutions designed and implemented by VLOPs with user safety as the main objective.

ALGORITHMS OF TRAUMA #2 – RISK ASSESSMENT INDEX CARD

3. Our case study is yet another anecdotal evidence confirming that. We do not know the exact actions of the participant in the past or outside of the scraping sessions and outside Facebook. We know however she has had a tendency to read about health-related issues online and over-focus on such information when confronted with it. It is therefore likely that she very strongly indicated, through her online behaviour, her interest and engagement in the problematic content, both in the past and maybe even in the course of our study (eg. by clicking on the health-related content), thus counteracting the dis-interest expressed through the “Hide post – See fewer posts like this” button. This suggests that signals collected based on behavioural observations had much more significant weight than the user’s explicit feedback which Facebook largely disregarded.

Observed negative effects	Excessive exposure to content aggravating mental health issues, which users feel they cannot escape (due to ineffectiveness of human agency tools allowing to flag unwanted content in order to limit its representation in the feed)
Categories of systemic risks corresponding to the observed negative effects (as in art. 34.1 DSA)	(b) negative effects for the exercise of fundamental rights (esp. right to privacy) (d) negative effects to the protection of public health and serious negative consequences to the person’s mental well-being
Factors indicating the risk is “systemic”	Observed negative effect is a consequence of how Meta chose to optimise its recommender system (to engage users’ attention at all costs and maximise time spent on the platform), which reflects the platform’s underlying business model. It has been established that negative user experience resulting from ineffectiveness of explicit feedback tools is not reserved to Facebook and may concern other VLOPs.
Factors influencing the systemic risks (as in art. 34.2 DSA)	(a) the design of their recommender systems and any other relevant algorithmic system

<p>Specific platforms' features to look at</p>	<ul style="list-style-type: none"> ▶ Signals that the algorithm is using to recommend content; ▶ Availability and effectiveness of human agency (explicit feedback) tools provided by the platform, in particular features allowing users to flag and de-rank specific types of content in their own feed; ▶ Terms and conditions explaining human agency tools made available by the platform and how they can influence the feed.
<p>Proposed examples of mitigation measures</p>	<ul style="list-style-type: none"> ▶ Introducing a filtering system aimed at filtering out content on topics flagged as undesired by the user from the list of suggestions made by the recommender systems (i.e. splitting the judgement of what would “engage” a user and filtering out what a user does not want to see). ▶ Alternatively: recalibrating ranking scores to increase impact of signals that reflect users’ explicit preferences (so that ranking scores of posts flagged as undesired and similar content decrease significantly). However in that case “do not show me posts like this” signal would still have to “compete” with the fact that user may be engaging with the topic (providing therefore “contradictory” signals to the platform). That is why introducing a filtering system, as suggested above, seems to be a more effective measure (as well as a more auditable one - see the next bullet point); ▶ Providing users with features allowing them to verify how their explicit feedback influenced selection of content that has been recommended to them. It could be for example the interactive tool “See how your feed has changed”, through which users can monitor whether their choices have been respected by the platform (Panoptykon et al., 2023 [2]⁴), ideally indicating how many/which posts were filtered out that would otherwise have been shown. ▶ Providing content curation features which encourage and allow users to pre-define their preferences about the type of content they wish to see or do not wish to see (in contrast to feedback tools, which users can only use after they saw unwanted recommendations).

4. Such a feature is particularly important in the context of the researchers’ findings that a lack of/unclear feedback provided to the user after clicking on the control tools discourages them from using those tools (Smith et al., 2021).

Conclusions. Safety buttons: a low hanging fruit to improve user experience

The results of our case study confirmed large amounts of problematic content in the participant's feed. They also suggest that the explicit feedback tool provided by the platform (the "Hide post – See fewer posts like this" button) proved ineffective (i.e. did not have the expected mitigating effect). Frequent and consistent use of this tool had no significant influence on the user's experience.

The case study adds to a pile of evidence indicating that recommender systems that prioritise user engagement over other metrics create a systemic risk by over-exposing users to harmful or toxic content. "Engagement" often comes at the price of exploiting their vulnerabilities and sensitive features, and incentivizes clicking and scrolling against their conscious intention not to look at certain content. We expect this risk to be recognised in VLOPs' risks assessments and adequately mitigated (solved in a systemic way).

Preventing social media recommender systems from pushing vulnerable users into dangerous "rabbit holes" (including those that aggravate mental health issues) is by no means an easy task, given that this effect is linked to the business model chosen by VLOPs. At the very least, however, VLOPs should provide users with tools ("safety buttons") allowing them to escape the vicious circle of exposure to what users themselves perceive as harmful content. More importantly: all those feedback and control features that are made available to users should work and bring expected mitigating effects. Providing "deceptive" features that, from a user perspective, do not have any positive impact on their experience, means – on the one hand – that the platform has not implemented adequate mitigation measures under art. 35 of the DSA, and – on the other hand – it also violates obligations imposed on online platforms by art. 27.3 in conj. with art. 25 of this Act.

It is important to, once again, stress that effective "safety buttons" do not entirely solve the problem of toxic feeds, which is far more complex and rooted in the business model of many VLOPs. It is clear that more needs to be done to re-engineer recommender systems away from engagement-based ranking and that, in principle, users' safety must be protected "by default" (also for users who do not use feedback tools)⁵. Therefore "safety buttons" should rather be seen as one of many mitigation measures necessary to address harmful effects related to the functioning of recommender systems.

Having said that, effective "safety buttons" seem to be a fairly "easy" measure to implement in order to protect users from at least one kind of harmful experience. In particular, this mitigation measure would still make a real difference for those who – like the participant of our case study – are flooded with legal but disturbing content in their personalized feeds.

Limitations

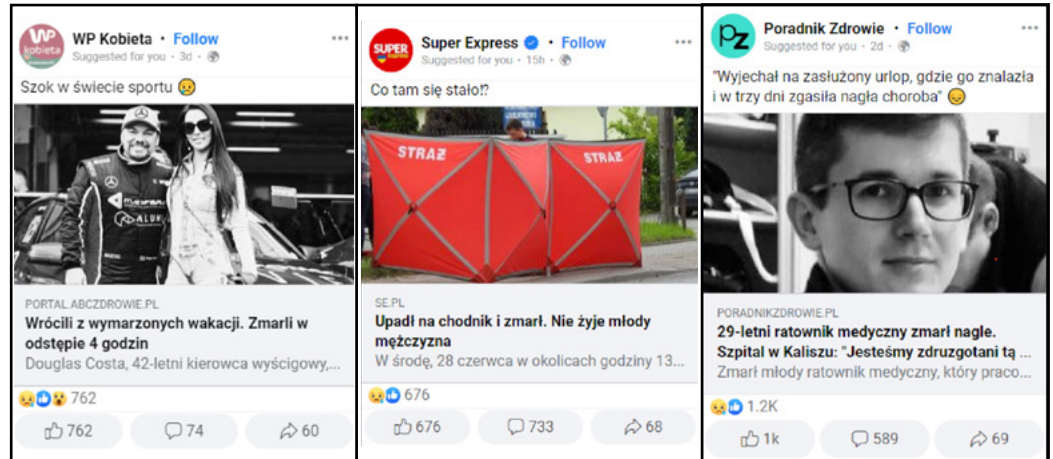
The authors note that this case study was only done on one person and the conclusions might not generalise to larger populations. Unfortunately, performing this kind of privacy-intensive research at scale is difficult, and often strongly discouraged by the VLOPs (Hatmaker, 2021). Nevertheless, the findings on personalisation are in line with other experiments performed at larger groups of users in the context of sponsored content (Ali et al., 2023).

5. "Users should not be responsible for making their experience on social media platforms safe. Safe defaults are paramount, considering that most users lack the awareness, time or skills to customise their experience. This should include top-down interventions to ranking algorithms, in order to make them less dependent on engagement, and thereby safer for all users." (Panoptikon et al., 2023 [2])

ANNEX

Examples of problematic “Suggested for you” content in the participant’s feed

THE SUDDEN DEATH OF YOUNG PEOPLE



“The sport’s world is in shock”

“They came back from a dream vacation. They died within 4 hours of each other”

“What happened?”

“He fell on the sidewalk and perished. A young man is dead”

“He went on a well-deserved vacation, where he was found and defeated by a sudden illness in three days”

“A 29-year-old paramedic died suddenly. Hospital in Kalisz: We are devastated...”

SEEMINGLY TRIVIAL SYMPTOMS OF TERMINAL DISEASES



“He had only one symptom. The doctors said it was anxiety.”

“The doctors sent him home. The next day he was dead”






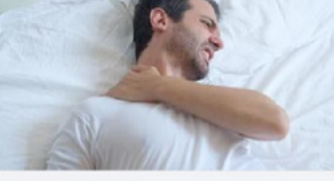
“Doctors said it's just a gut problem”

“Her symptoms intensified after she had given birth. This is how cancer was disguised”



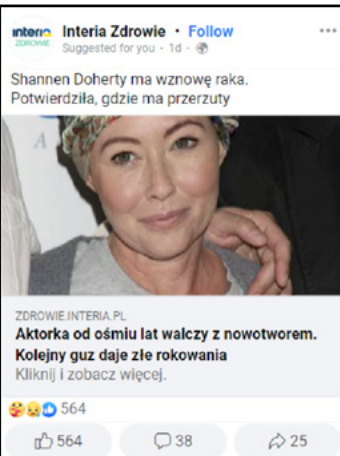

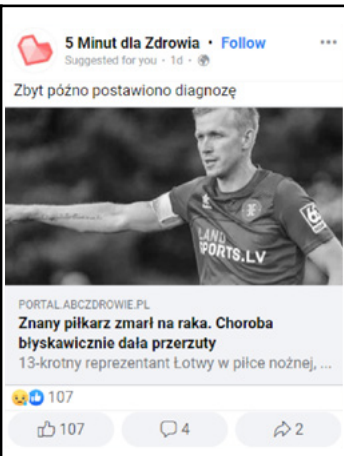

“The GP thought the man had a muscle spasm.”

“Shoulder pain turned out to be a symptom of pancreatic cancer. The 54-year-old died three months later”

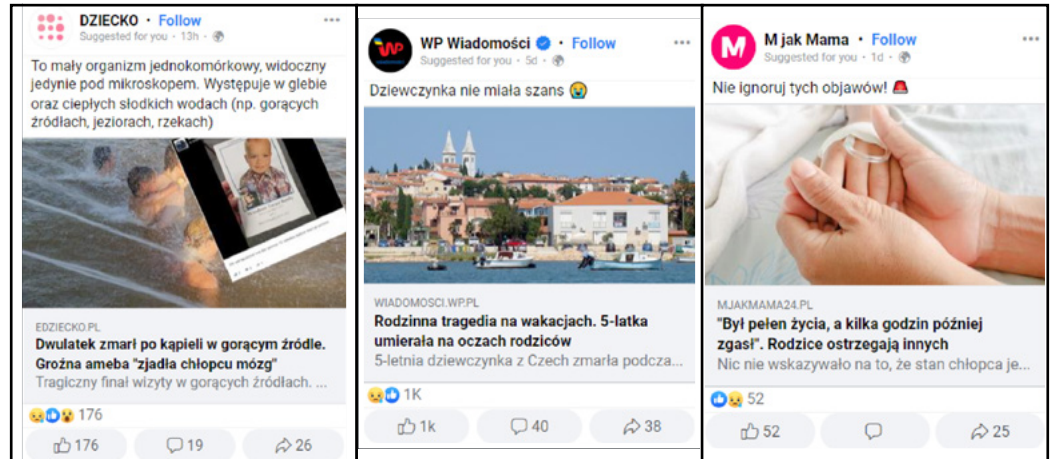
ALARMING
PHYSICAL SIGNS

 <p>Poradnik Zdrowie · Follow Suggested for you · 4d · 🌐</p> <p>Gdy je zauważysz, nie zwlekaj z wizytą u lekarza. 🙏🙏</p>  <p>PORADNIKZDROWIE.PL Te objawy raka wątroby łatwo pomylić z niestrawnością. Nie ignoruj ich W większości przypadków objawy raka wąt...</p> <p>👍 69 💬 4 ➦ 43</p>	 <p>interia Zdrowie · Follow Suggested for you · 6h · 🌐</p> <p>Objawy, które mogą być pierwszymi sygnałami raka. Nie budzą podejrzeń 🙏</p>  <p>ZDROWIE.INTERIA.PL Pięć objawów, które mogą być pierwszymi sygnałami rozwijającego się raka Kliknij i zobacz więcej.</p> <p>👍 16 💬 3 ➦ 13</p>	 <p>WP abcZdrowie.pl · Follow Suggested for you · 3d · 🌐</p> <p>Ważny sygnał wysłany przez organizm</p>  <p>PORTAL.ABCZDROWIE.PL Ból, który może zwiastować tętniaka. Uważaj to tykająca bomba Na bóle kręgosłupa skarżą się często już mło...</p> <p>👍 104 💬 32 ➦ 76</p>
<p>“If you notice them, do not delay visiting a doctor.”</p> <p>“These symptoms of liver cancer can be easily confused with indigestion. Don't ignore them.”</p>	<p>“Symptoms that may be the first signs of cancer. Usually they do not seem suspicious”</p> <p>“Five symptoms that may be the first signs of developing cancer”</p>	<p>“An important signal sent by the body”</p> <p>“This ache may be a sign of an aneurysm. Watch out, it’s a ticking time bomb”</p>

THE ILLNESS AND
DEATH OF CELEBRITIES

 <p>WP Wirtualna Polska · Follow Suggested for you · 2d · 🌐</p> <p>Tajemnicze zaginięcie w końcu ma swój smutny finał 😞</p>  <p>FILM.WP.PL Julian Sands nie żyje. Na szczątki zaginionego aktora natrafili turyści Tajemnicze zaginięcie Juliana Sandsa w koń...</p> <p>👍 72 💬 19 ➦</p>	 <p>interia Zdrowie · Follow Suggested for you · 1d · 🌐</p> <p>Shannen Doherty ma wznówę raka. Potwierdziła, gdzie ma przerzuty</p>  <p>ZDROWIE.INTERIA.PL Aktorka od ośmiu lat walczy z nowotworem. Kolejny guz daje złe rokowania Kliknij i zobacz więcej.</p> <p>👍 564 💬 38 ➦ 25</p>	 <p>5 Minut dla Zdrowia · Follow Suggested for you · 1d · 🌐</p> <p>Zbyt późno postawiono diagnozę</p>  <p>PORTAL.ABCZDROWIE.PL Znany piłkarz zmarł na raka. Choroba błyskawicznie dała przerzuty 13-krotny reprezentant Łotwy w piłce nożnej, ...</p> <p>👍 107 💬 4 ➦ 2</p>
<p>“The mysterious disappearance finally has its sad ending”</p> <p>“Julian Sands is dead. The remains of the missing actor were found by tourists”</p>	<p>“Shannen Doherty's cancer has returned. She confirmed where the metastases are”</p> <p>“The actress has been battling cancer for eight years. Another tumor means a grim prognosis”</p>	<p>“The diagnosis came too late”</p> <p>“Famous football player dies of cancer. The disease spread like wildfire”</p>

ACCIDENTS AND DEATHS INVOLVING YOUNG CHILDREN



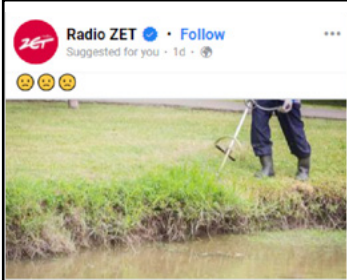
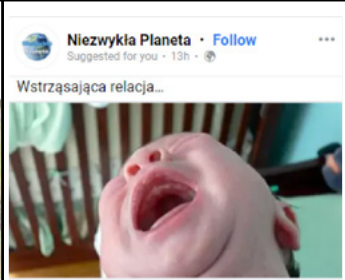

<p>“A two-year-old died after bathing in a hot spring. A dangerous amoeba 'ate the child's brain”</p>	<p>“The girl didn't stand a chance” “Family tragedy strikes on vacation. The 5-year-old died in front of her parents' eyes was dying in front of her parents.”</p>	<p>“Do not ignore these symptoms.” “He was full of life and died a few hours later. His parents warn others.”</p>
---	--	---

DRAMATIC ACCIDENTS AND MURDERS



<p>“5 meters from the road there is a dead body of a young woman” “Her last messages to a friend were ‘Help me’, ‘They will kill me’”</p>	<p>“Macabre” “They were burned alive tied to hospital beds. A nightmare in Częstochowa”</p>	<p>“They went to the hospital for help and were given their child back in a white coffin” “Horror in the hospital. 6-year-old girl crushed in an elevator!”</p>
---	---	---

MORBID STORIES

 <p>Radio ZET • Follow Suggested for you · 1d · 🌐</p> <p>☹️☹️☹️</p> <p>WIADOMOSCI.RADIOZET.PL Tragedia podczas koszenia trawy. Mężczyzna odciął sobie jądra 39-letni Pradistin Chuipad odciął sobie genit...</p> <p>👍👍👍 1.7K 👍 1k 💬 801 ➦ 222</p>	 <p>Niezwykła Planeta • Follow Suggested for you · 13h · 🌐</p> <p>Wstrząsająca relacja...</p> <p>FANTUBE.PL Zapłakana matka próbowała połączyć główkę dziecka z jego ciałem. Tragiczne w... Krążący w tym czasie w Internecie artykuł zo...</p> <p>👍👍👍 9 👍 9 💬 2 ➦ 1</p>	 <p>SUPER Super Express • Follow Suggested for you · 1d · 🌐</p> <p>Co za koszmar 😱</p> <p>SE.PL Kosiarka do trawy zabiła piękną 27-latkę! Rodzina znalazła jej zęby Horror w parku! Pracownik koszący trawę urz...</p> <p>👍👍👍 603 👍 603 💬 164 ➦ 37</p>
<p>“Man cuts off own balls in lawn-mower accident”</p>	<p>“Shocking...”</p> <p>“The crying mother attempted to attach the child’s head back to the body”</p>	<p>“What a nightmare”</p> <p>“A lawn-mower killed a beautiful 27-year-old. Her family found her teeth.”</p>

Ali, M., Goetzen, A., Mislove, A., Redmiles, E. M., & Sapiezynski, P. (2023), *Problematic Advertising and its Disparate Exposure on Facebook* (arXiv:2306.06052), arXiv, <https://doi.org/10.48550/arXiv.2306.06052>

Hatmaker T. (2021), *Facebook cuts off NYU researcher access, prompting rebuke from lawmakers*, TechCrunch, <https://techcrunch.com/2021/08/04/facebook-ad-observatory-nyu-researchers/>

Integrity Institute (2023), *Resource on Platform Risk Assessments and Compliance* (Draft)

Jin Y., Cardoso B. & Verbert K. (2017), *How Do Different Levels of User Control Affect Cognitive Load and Acceptance of Recommendations?*, https://www.researchgate.net/publication/318778836_How_Do_Different_Levels_of_User_Control_Affect_Cognitive_Load_and_Acceptance_of_Recommendations

Liu A., Wu S. & Resnick P. (2023), *How to Train Your YouTube Recommender to Avoid Unwanted Videos* (arXiv:2307.14551v2), arXiv, <https://doi.org/10.48550/arXiv.2307.14551>

Meta, *Learn about and manage suggested content in your Facebook Feed* (viewed: 1 December 2023), Help Center, https://www.facebook.com/help/485502912850153/?helpref=related_articles

Milton A., Ajmani L., DeVito M.A. & S. Chancellor (2023), *I See Me Here: Mental Health Content, Community, and Algorithmic Curation on TikTok*, CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, <https://dl.acm.org/doi/10.1145/3544548.3581489>

Mozilla Foundation (2022), *Does this button work? Investigating YouTube's ineffective user controls*, <https://foundation.mozilla.org/en/research/library/user-controls/report/>

Panoptikon (2021), *Algorithms of trauma: new case study shows that Facebook doesn't give users real control over disturbing surveillance ads*, <https://en.panoptikon.org/algorithms-of-trauma>

Panoptikon, Irish Council for Civil Liberties & People vs Big Tech (2023) [1], *Fixing Recommender Systems. From identification of risk factors to meaningful transparency and mitigation*, https://panoptikon.org/sites/default/files/2023-08/Panoptikon_ICCL_PvsBT_Fixing-recommender-systems_Aug%202023.pdf

Panoptikon & People vs Big Tech (2023) [2], *Prototyping user empowerment: Towards DSA-compliant recommender systems*, https://panoptikon.org/sites/default/files/2023-11/peoplesvsbigtech_panoptikon_prototyping-empowerment_brief_20112023.pdf

Smith K., Bullen G., & Huerta M. (2021), *Dark Patterns in User Controls: Exploring YouTube's Recommendation Settings*, <https://simplysecure.org/blog/dark-patterns-in-user-controls-exploring-youtubes-recommendation-settings/>

White, R. W. & Horvitz, E. (2009), *Cyberchondria: Studies of the escalation of medical concerns in Web search*, <http://doi.acm.org/10.1145/1629096.1629101>