

Users affected by Facebook suspensions and content deletions in CEE countries:

Quantitative research results

Péter Náadori*, Árpád Rab**

ABSTRACT

Facebook users' speech and the company's moderation activities are increasingly perceived as central to – and, often, detrimental to – the public sphere of several societies. Academics, policy-makers and concerned citizens, though, are often short of empirical data to base related arguments, proposals and/or apocalyptic visions on. A representative survey of the adult population of four Central and Eastern European (CEE) countries – Czechia, Hungary, Poland and Romania – by Hungary's Institute of the Information Society (ITKI) at the University of Public Service, concerning questions of "Trust, Awareness and Alarm on the Internet", produced such data on certain topics.¹ We learned how many people used Facebook, how many of them were affected by the company's moderation decisions – user suspensions and content deletions –, how many of them appealed, and how the affected users perceived and evaluated these decisions. This paper presents the relevant results of the research and outlines some possible – maybe preferable – directions for further work.

KEY FINDINGS

- In the four surveyed countries, 67-85% of adult internet users said they were using Facebook at least once a month. (No other social media platform was remotely that popular).
- Bans and deletions were affecting 9,5-14% of Facebook users in the surveyed countries.
- 33-47% of affected users said that they had not been informed about the reason of the ban/deletion at all.
- 16,5-36,5% of the affected users reported that they had appealed Facebook's decision, 30-57% of them successfully.

KEYWORDS: social media moderation, platform governance, Facebook studies

* Project consultant, ITKI.

** Lead researcher, ITKI.

¹ "Trust, Awareness and Alarm on the Internet", Institute of the Information Society, 15 July 2020, <https://iis.uni-nke.hu/hirek/2020/07/15/trust-awareness-and-alarm-on-the-internet>.

1. Introduction

Regarding the issues of content moderation and user governance, Facebook is often criticized for not doing enough. To cite one of the more recent and prominent examples, the summer of 2020 saw what a news report labeled “the largest advertiser boycott in the site’s history”.² The boycott, organized by the Anti-Defamation League-affiliated *Stop Hate for Profit* campaign, aimed to pressure the company – and founder/CEO Mark Zuckerberg personally – to take “meaningful action” against “hate, bigotry, racism, antisemitism, and disinformation.”³ The campaign expressly called for Facebook to “[f]ind and remove public and private groups focused on white supremacy, militia, antisemitism, violent conspiracies, Holocaust denialism, vaccine misinformation, and climate denialism.”⁴

Regarding the issues of content moderation and platform governance, Facebook is often criticized for doing too much. About a month before the advertiser boycott, US president Donald J. Trump issued his *Executive Order on Preventing Online Censorship*. It was based on the premise that online platforms – including Twitter, seemingly the president’s immediate target, but Facebook as well – abused their “immense, if not unprecedented, power to shape the interpretation of public events; to censor, delete, or disappear information; and to control what people see or do not see.” According to the order, “[t]ens of thousands of Americans have reported, among other troubling behaviors, online platforms »flagging« content as inappropriate, even though it does not violate any stated terms of service; [...] and deleting content and entire accounts with no warning, no rationale, and no recourse.” The order ostensibly set in motion federal processes, including the application of existing law and the proposition of new regulation, to end platforms’ “engage[ment] in deceptive or pretextual actions [...] to stifle viewpoints with which they disagree.”⁵

In early August, partaking in the efforts of the country’s Digital Freedom Working Group (chaired by the Minister of Justice), Attila Péterfalvi, head of Hungary’s National Authority for Data Protection and Freedom of Information proposed new regulation that “would make it Hungary’s competence to review the suspensions of personal [Facebook] profiles”. Quoted in a media report, he said he had recommended the establishment of a new type of

² Alex Hern, “Facebook Moderators Call for Advertiser Boycott to Be Extended”, The Guardian, 3 August 2020, <https://www.theguardian.com/technology/2020/aug/03/facebook-moderators-call-for-advertiser-boycott-to-be-extended>.

³ “Dear Mark”, Stop Hate for Profit, accessed 3 August 2020, <https://www.stophateforprofit.org>.

⁴ “Recommended Next Steps”, Stop Hate for Profit, accessed 3 August 2020, <https://www.stophateforprofit.org>.

⁵ “Executive Order on Preventing Online Censorship” (2020), <https://www.whitehouse.gov/presidential-actions/executive-order-preventing-online-censorship/>.

proceeding of the authority, “during which the authority, to positively preserve freedom of opinion, orders Facebook to review the gratuitous suspensions.”⁶

These are the tips of some extremely large and jagged icebergs set towards each other on a collision course. In the last 10-15 years countless articles, books, comments, posts, speeches and tweets have been formulating these remonstrations, along with many other criticisms of Facebook and other platforms, and there is no immediate agreement, compromise or solution in sight. (The company maintains that it allocates enormous resources to the fight against hate speech, and denies that it would be anything but viewpoint-neutral vis-à-vis any speaker who adheres to its Community Standards.⁷)

We consider this research a small but substantial step towards a better understanding of what happens between Facebook and its users, hopefully a starting point for more extensive and increasingly more appreciative studies.

2. Methodology

The data collection was carried out by the MASMI European Market Research Network in November 2019. Data are representative in each country by age, gender, education, type of settlement and region. The sample sizes were as follows: Hungary 1003, Czech Republic 1004, Poland 1000, Romania 1008.

3. Facebook usage

Internet penetration (defined as using the internet regardless of mode and place of access) was reported to be between 70% and 80% in the four countries. Facebook was found to be one of the most popular online services/activities: the lowest percentage of internet users who said they used the site at least once a month was still 67,4 (Poland), the highest value being 85,7% (Romania). (See *Figure 1*.)

⁶ Orsolya Nagy, "Sérti a véleménynyilvánítás szabadságát a Facebook indokolatlan cenzúrája", Magyar Nemzet, 1 August 2020, <https://magyarnemzet.hu/belfold/velemenynyilvanitast-sert-a-facebook-indokolatlan-cenzuraja-8451747/>.

⁷ "Community Standards", Facebook, accessed 3 August 2020, <https://www.facebook.com/communitystandards/>.

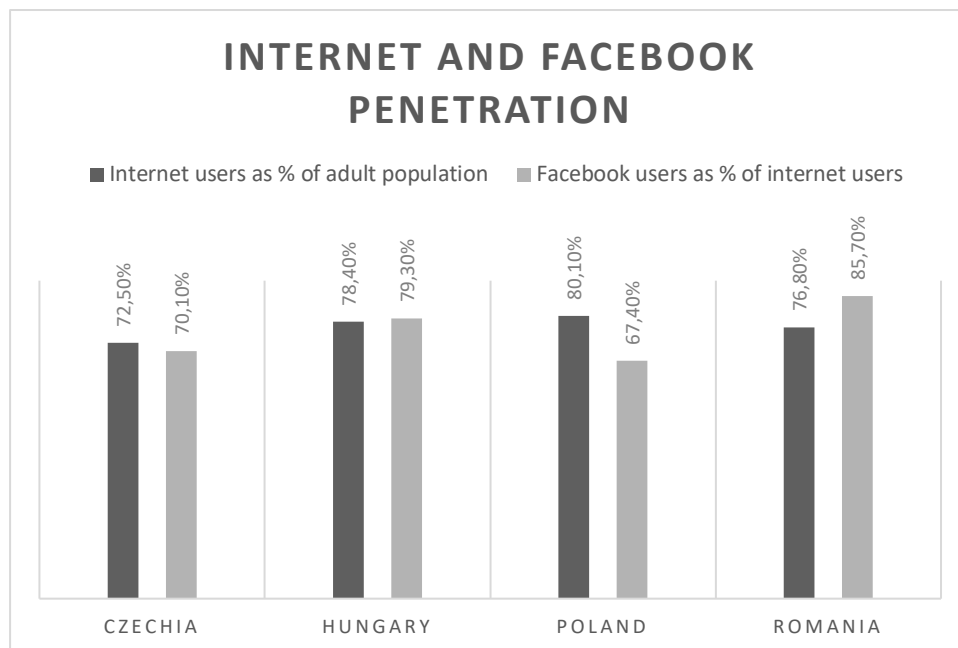


Figure 1

The share of Facebook users who reported that they used the site “monthly or less frequently” was low: 3% in Czechia, 3,8% in Hungary, 2,4% in Poland, and 1,1% in Romania. Heavy users were shown to be the relative or absolute majority everywhere, with 36,7% (Czechia), 58,9% (Hungary), 48,2% (Poland) and 67,5% (Romania) going to Facebook at least once a day.

4. Users’ experience with suspensions and deletions

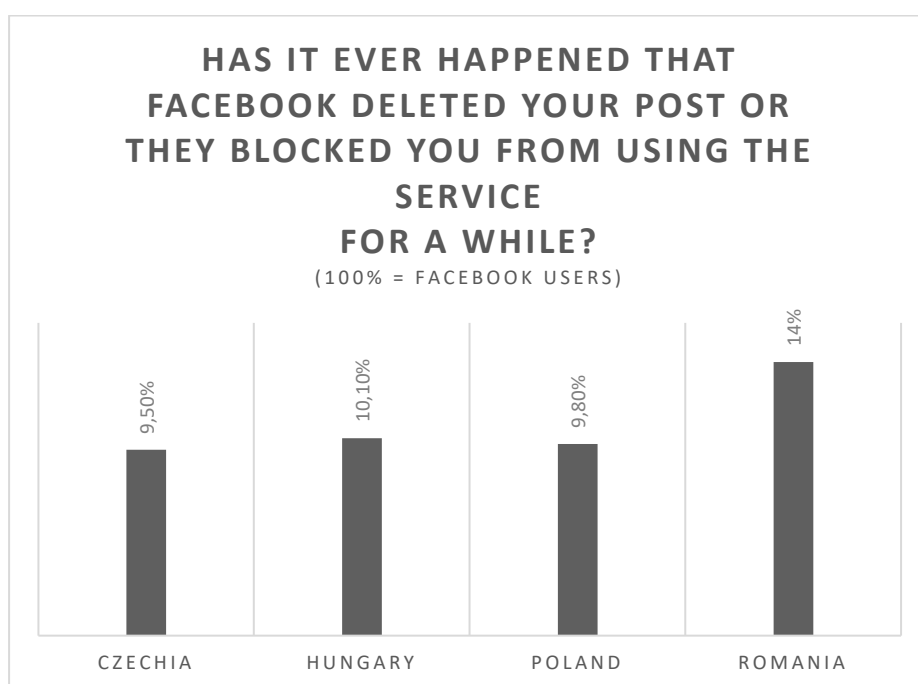


Figure 2

The unweighted average of users who claimed to have been affected by suspensions and/or post deletions on Facebook was 10,85% across the four countries. As *Figure 2* above shows, the percentage of Czech, Hungarian and Polish Facebook users who reported to be affected was rather consistent at about 10%. The relative outlier here was Romania with its higher number, 14%.

Across the four countries, at least a third of affected users said that they had not been informed of the reason their content had been deleted or their account suspended. The percentage of those saying they've been given a reason and have been sufficiently satisfied by that showed strong divergence: only 16,8% in Czechia, while as many as 45,10% in Romania. In all of Czechia, Hungary and Poland, roughly a third of affected users said that they had received some communication from Facebook regarding the decisions, but had found the communicated reasons wanting. (See *Figure 3*.)

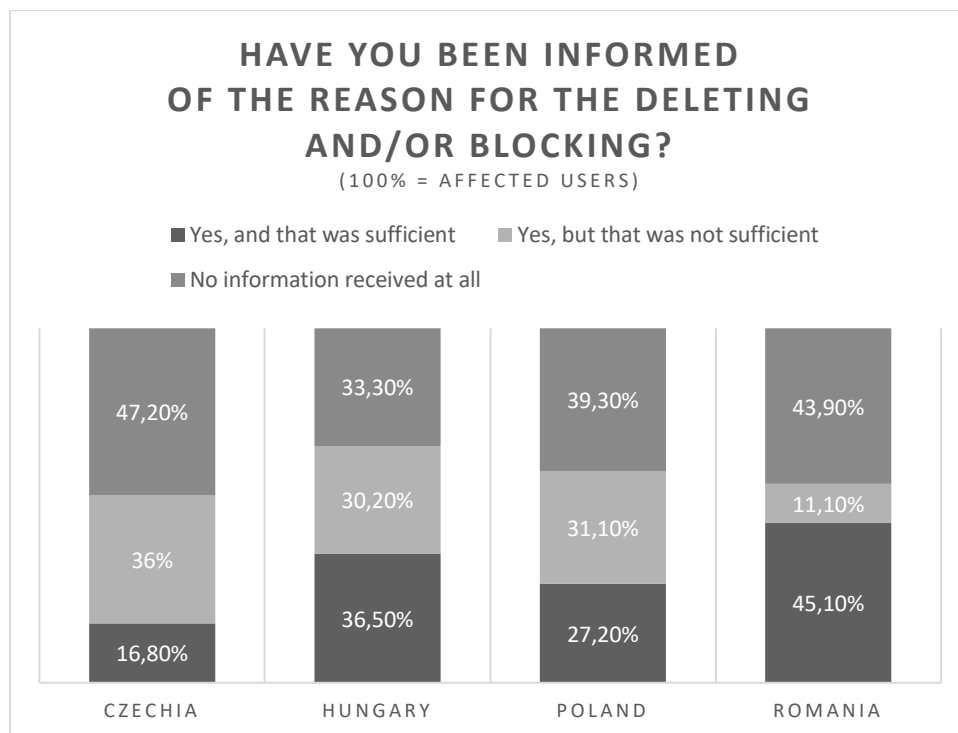


Figure 3

It must be stressed that the data here are reflecting users' perceptions and self-constructed narrative and do not say anything about whether the reported grievances were founded or unfounded, reasonable or unreasonable. It is perfectly conceivable, for example, that some or even many of those who reported to be kept in the dark had, in fact, simply forgot about some notices from Facebook by the time they answered the survey's questions.

In certain cases, Facebook makes it possible for users to ‘appeal’ the deletion of their content. (According to 2019 report of the Data Transparency Advisory Group set up by Facebook, “[w]hen a user elects review, a different human content reviewer takes a second look at the post. The second reviewer is at the same level as the first one; they are not a higher-level content reviewer with more training or expertise. In this sense the term appeal may be misleading. Review is perhaps more accurate.”⁸) In Hungary, Poland and Romania, about a quarter to a third of affected users said that they appealed the moderation decisions. In Czechia, only 16,6% reported doing so. (See *Figure 4*.)

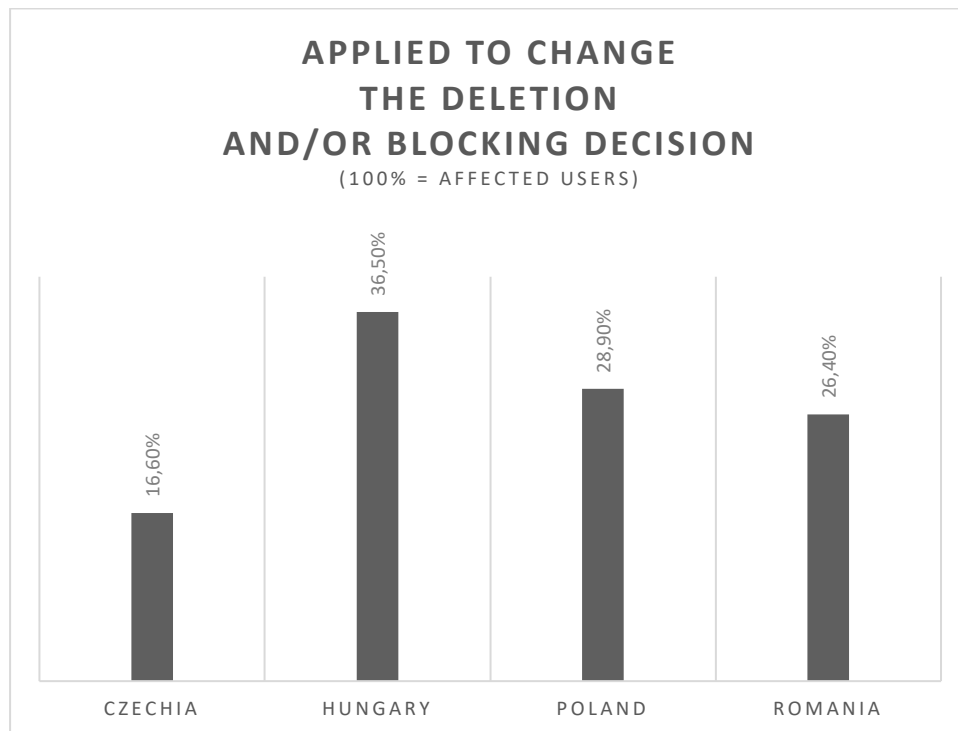


Figure 4

According to these users, Facebook overturned at least 30% of the appealed decisions in all four of the surveyed countries, with the percentage going as high as 57% in Poland. (See *Figure 5*.) This means that out of all users affected by deletions and/or suspensions, 7,1% (Czechia), 12% (Hungary), 16,5% (Poland) and 10,8% (Romania) reported that they experienced Facebook overturning its decisions on appeal.

⁸ Ben Bradford és mtsai., "Report of the Facebook Data Transparency Advisory Group" (New Haven: The Justice Collaboratory, Yale Law School, 2019), 14.

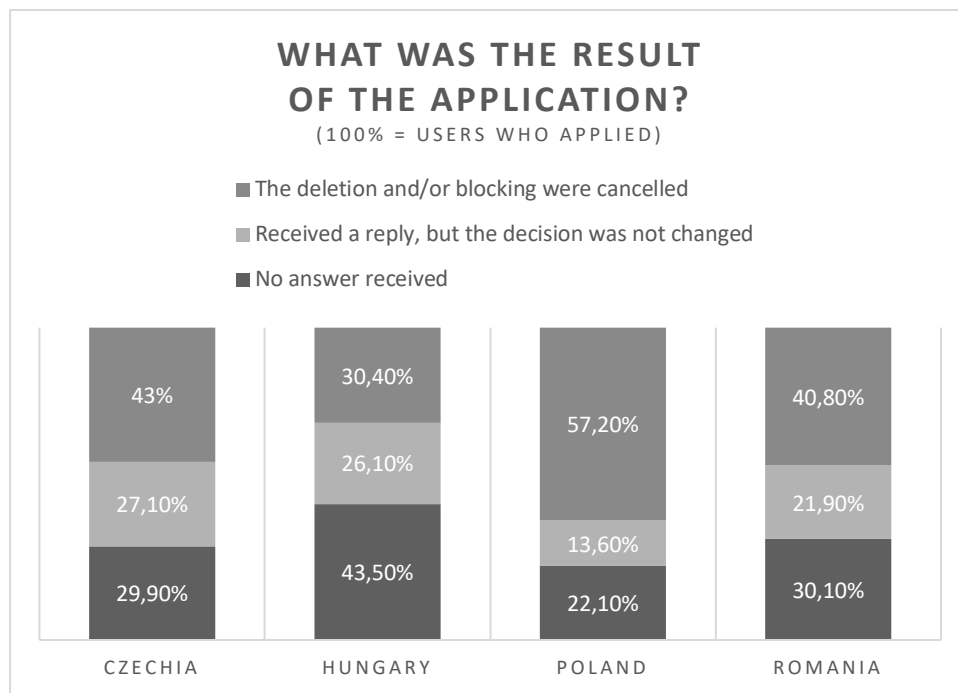


Figure 5

5. Further research: content focus and user focus

Regardless of the above-mentioned criticisms coming from all directions, and regardless of the company's supposed or actual motivations, it can be said that, over the years, Facebook has been moving toward greater transparency in terms of content moderation.⁹ The company made its internal moderation guidelines public in 2018.¹⁰ In the same year, Facebook started the regular publication of its so-called Community Standards Enforcement Report (CSER), "providing metrics on how well [they] enforced [their] policies".¹¹

A defining characteristic of the CSER so far is that it mostly presents metrics based on the amount of *pieces of content* involved. It provides, by category (Adult Nudity and Sexual Activity, Bullying and Harassment, Child Nudity and Sexual Exploitation of Children, Dangerous Organizations: Terrorism and Organized Hate, Hate Speech, Regulated Goods: Drugs and Firearms, Suicide and Self-Injury, Violent and Graphic Content), the quarterly amount of 'actioned' (deleted) posts, the amount of posts flagged by users, the amount of posts flagged proactively by Facebook, the amount of content appealed, and the amounts restored without or after an appeal. As per the CSER, for example, in the 4th quarter of 2019, Facebook deleted 120,7 million pieces of content breaching the Community Standards in

⁹ Chris Sonderby, "Our Continuing Commitment to Transparency", *About Facebook* (blog), 2020. május 12., <https://about.fb.com/news/2020/05/transparency-report/>.

¹⁰ Monika Bickert, "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process", *About Facebook* (blog), 2018. április 24., <https://about.fb.com/news/2018/04/comprehensive-community-standards/>.

¹¹ Guy Rosen, "Community Standards Enforcement Report, May 2020 Edition", *About Facebook* (blog), 2020. május 12., <https://about.fb.com/news/2020/05/community-standards-enforcement-report-may-2020/>.

these categories. About 4,9% – almost 6 million individual pieces of content – went into the appeal process. About 1 million pieces of content got restored this way, meaning that 16,7% of the appeals were successful. All in all, less than 1% of the deleted pieces of content have been eventually de-erased.¹²

While the CSER is a highly commendable effort of transparency, and a treasure trove of data, focusing on individual pieces of content is an approach that has limitations. In practice, platform governance is not merely the policing of individual posts. Facebook is actively “penalizing borderline content so it gets less distribution and engagement.”¹³ It also suspends user accounts for various periods of time – ostensibly depending on the severity of the offence and/or the affected user’s history –, or deletes them permanently “for severe violations of the Facebook Community Standards.”¹⁴ As Gillespie noted, “[t]he experience of [...] content moderation is at an *intimate scale*, even if that user knows, intellectually, that moderation is an industrial-sized effort. [»T]hat’s my post you deleted; my account you suspended. What did I do that was so wrong?»”¹⁵

Without user-focused metrics – like the number and characteristics of users who breach the Community Standards in different categories, the effectiveness of deletions/suspensions in persuading users to toe the line etc. – and survey-based data on user attitudes and perceptions, our understanding of the field cannot be anything but incomplete. It’s worth noting in the European Commission’s latest report based on the monitoring of platform adherence to the *Code of conduct on countering illegal hate speech online*, Facebook have been praised as “the only platform that provides systematic feedback to almost all users” “on the outcome of their notifications”.¹⁶ As articles by Myers West and Jhaver et al. attest, even studies with smaller samples and narrower foci can be enlightening.¹⁷

The data presented here are based on just four questions of a much wider-ranging questionnaire, and as such, the findings obviously lack nuance and complexity. They are, however, fit for the purpose of providing a clear indication of the scale and some basic

¹² “Community Standards Enforcement”, accessed 12 September 2020, <https://transparency.facebook.com/community-standards-enforcement>.

¹³ Mark Zuckerberg, “A Blueprint for Content Governance and Enforcement”, *Facebook* (blog), 15 November 2018, <https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/>.

¹⁴ “Disabled Accounts”, Facebook Help Center, accessed 21 September 2020, <https://www.facebook.com/help/185747581553788>.

¹⁵ Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (New Haven & London: Yale University Press, 2018), 141–42.

¹⁶ European Commission - Questions and answers, “The Code of Conduct on Countering Illegal Hate Speech Online”, Text, European Commission, 20 June 2020., https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1135.

¹⁷ Sarah Myers West, “Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms”, *New Media & Society* 20, sz. 11 (2018): 4366–83; Shagun Jhaver et al., “Did You Suspect the Post Would Be Removed?”, *Proceedings of the ACM on Human-Computer Interaction*, 2019., <https://dl.acm.org/doi/abs/10.1145/3359294>.

features of the issues addressed. The survey – or an extended version of it – should be repeated at other dates and in other countries to supply longitudinal and comparative data.

More detailed surveys should cover

- the type/types of moderation a user have encountered,
- the time, frequency and extent of deletions and/or suspensions,
- the characteristics and circumstances of the content/conduct that led to enforcement action/actions,
- Facebook's reasoning (if provided),
- characteristics of the affected user segments (demographics, value systems, political affiliation etc.),
- users' perception regarding the cause of moderation (i.e. their own fault, being uninformed about standards, political bias, other users' malice etc.),
- the appeals process,
- user knowledge and attitudes about Facebook's Community Standards,
- users' perception regarding the decisions being made by algorithms and/or humans.

As discussed, more than 10% of all Facebook users in the four surveyed CEE countries have been affected by moderation decisions, and, on unweighted average, less than a third of them reported to have been given sufficient reasons. Dedicated follow-up research should be designed to provide further insight into the realities and dynamics of Facebook's moderation activities vis-à-vis users.