The Case of a Facebook Content Moderation Debacle in Greece

Charis Papaevangelou, Nikos Smyrnaios

A Facebook content moderation debacle: the case of Greece

Charis Papaevangelou, PhD candidate, LERASS, University of Toulouse
Nikos Smyrnaios, Associate professor, LERASS, University of Toulouse

(Draft of a chapter to be published in Jebril N., Iordanidou S. (ed .), *Digital Journalism in Emerging Media Markets*, Palgrave Macmillan)

In the first months of 2021, a story gained prominence and public attention in Greece: one involving Facebook's obscure content moderation system and Dimitris Koufontinas, a detainee who is serving life sentence for crimes related to domestic terrorism in Greece perpetrated as a member of the "Revolutionary Organization 17 November." Based on this particular case, in this chapter, we will demonstrate how the so-called "Big Tech" platforms' strategies, Facebook's in this case, may influence smaller or emerging media markets, like that of Greece, without much regard for socio-political implications or accountability. Our research aims to contribute to the multifaceted question of who governs the contemporary digital public sphere (Boeder, 2005), as well as to explore the political stakes of platforms' content moderation policies that, in some cases, may amount to censorship (Gillespie, 2018). Indeed, the functioning of democratic societies depends on a fine and complex trade-off between, on the one hand, the ability to express one's opinions freely in political debate and, on the other hand, the protection of citizens from abuses such as harassment, hate speech, misleading propaganda and political manipulation.

This chapter first discusses how Facebook's content moderation process work, along with Facebook's policies on Dangerous Individuals and Organizations, which were allegedly violated in the case we examine; second, we will briefly describe necessary contextual information regarding 17 November and Koufontinas's actions, along with what sparked the examined controversy. Finally, we will present and discuss several instances of content and account restrictions that took place during the same time span, mostly between February and April 2021, and that referred to Koufontinas. Last, we will conduct a case-by-case analysis to deduce if the content had indeed violated any platform policy in order to extrapolate what may have happened in this case and what socio-political stakes exist with the current content governance (Suzor, 2020).

How does Facebook's Content Moderation work?

Facebook has admittedly the largest and most sophisticated content moderation system in place compared to other major social media companies (Singh, 2019). This can be explained primarily due to its vast popularity; it still is the most popular social network with 2.8 billion active users, with YouTube following with 2.3 billion active users (Statista, 2021). Having to deal with so many users, tech companies have resorted to tech solutions to police content on their platforms at scale (Gillespie, 2020; Suzor, 2020); hence many of them are increasingly relying on automated solutions based on Artificial Intelligence (AI) and Machine Learning (ML) algorithms with dire consequences, in many cases, on human rights (Kassem & Marwa, 2021; Zuiderveen Borgesius, 2020). According to estimates, the digital content moderation industry is projected to reach $8.8 billion in 2022, roughly double the 2020 total. Facebook, in particular, is the biggest client of companies that undertake outsourced content moderation contracts, such as the multinational firm Accenture, whose contract with Facebook is worth at

least $500 million a year[1]. More than 15,000 moderators, out of a total of 200,000 worldwide, work for Facebook contractors.

In many countries Facebook plays a vital role in a society's news diet; for instance, in Greece, where trust in media has plummeted, Facebook is the most popular social platform with 57% of internet users selecting it for their news consumption (Newman et al., 2020, p. 72). Its popularity, along with its history of failures to stop campaigns of disinformation and toxicity, is also one of the reasons that Facebook is under the most scrutiny from the press and regulators (Gillespie, 2018). A recurrent subject of criticism is its content moderation system that is allegedly producing 300,000 wrong decisions every day[2].

Facebook's moderation can be either *ex-ante* or *ex-post*, that is either before or after a piece of content is published (Klonick, 2018, p. 1635). The *ex-ante* approach relies on automated solutions that screen the content that is being uploaded before it is published; for instance, there is a hashed database of known bad-faith actors that is shared among members of the Global Internet Forum to Counter Terrorism (GIFCT), a non-governmental organization that Microsoft, Twitter, Facebook, and YouTube created in 2017 to prevent terrorists from using their services. The content promoted by these actors is deleted before it is even published. Additionally, moderation can be either reactive, where "moderators passively assess content" or proactive, where "moderators actively seek out published content for removal" (Klonick, 2018, p. 1635). To this end, Facebook uses a combination of tens of thousands of human reviewers (Newton, 2021), who in many cases work in poor and dehumanizing conditions (Roberts, 2019), and opaque algorithmic processes to police content on its platform (Singh, 2019, p. 22).

As part of pipelining the process, Facebook has created a set of policies, called Community Standards, that its users must adhere to and that its moderators must follow to do their job. Yet, we know from leaked reports that there are serious problems with these standards. For instance, there are secret guidelines that Facebook provides its reviewers with (Hern, 2021) that are inaccessible to its users, as well as a whitelist that exempts quasi-VIP users from being held to the same standards (Horwitz, 2021).

Thanks to whistle-blowers and investigative reporting, we are now aware that Facebook employs three tiers of moderators: Tier 3 moderators do the majority of "day-to-day" reviewing, such as the vast majority of user reports, and are usually outsourced to third-party contractors outside the US, like multinational corporation Teleperformance in Greece (Παπαδόπουλος, 2019); Tier 2 moderators supervise Tier 3 moderators and are responsible for reviewing "prioritized or escalated content"; last, Tier 1 moderators are those who shape Facebook's policies and are based within its headquarters (Klonick, 2018, p. 1640-41).

In addition to these practices that further obscure Facebook's content moderation system and seed distrust among users, the platform does not divulge much information concerning the way that user reporting is handled. We know that the reported content awaits review by a Tier 3 human moderator; in the meantime, that content, having passed the automated filters, will stay intact. It is then up to the reviewer to either take it down, leave it up or enforce other

[1] https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html.
[2] https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=78ee731754d0.

actions like visibility restrictions; if unable to determine what should be done, in most cases, it will be escalated and assigned to a Tier 2 moderator. It remains unclear if there are specific policy violations that are prioritized over others or if there are types of content that will be "frozen" (i.e., restricted) until a verdict has been made.

The Case of Dimitris Koufontinas and 17 November

Before we begin discussing our case study, we ought to provide some necessary contextual information. Revolutionary Organization 17 November ("17 November" for short) was active between 1975 and 2002; their victims included Greeks and foreigners: former torturers of Greece's seven year-long junta between 1967-1973, businesspeople, publishers, journalists (including the brother-in-law of current Prime Minister Kyriakos Mitsotakis), judges, police officers, secret services members, officials and ordinary staff of the US, UK and Turkish embassies in Athens, and in one instance a bystander, student Thanos Axarlian. The perpetrators were arrested in 2002 and put on trial. Several 17 November members, including Koufontinas, are still serving long prison sentences. In addition, the United States" Secretary of State had designated the organization as a Foreign Terrorist Organizations in 10/8/1997 and delisted them in 9/3/2015[3].

After New Democracy, the traditional conservative political party of Greece, won the elections in 2019, one of the first things they did was to amend the Penal Code, which they were relentlessly attacking as the country's major opposition for being too lenient on "terrorists." Among the passed amendments, one was specifically aimed at Koufontinas: convicted terrorists would have the right to apply for parole only after having served 22 years of their sentence, and not 17 as was previously foreseen (Μάνδρου, 2019); Koufontinas had served 16 at the time. A few months later, the government passed another bill that abolished the right for convicted terrorists to serve at the country's rural prisons, which are thought to have better living conditions, and had to be instead transferred to the previous serving facility where they were held.

Up until 2018, Koufontinas was serving at the rural prison of Kassavetia Volou; prior to that, he was being held in the country's largest and most secure prison complex near Athens, Korydallos. So, with the new bill, Koufontinas had to return there. However, during the same time, the government announced it would reform the Korydallos detention center to a prison only for detainees awaiting trial; thus, it was decided to transfer Koufontinas to Domokos, 220 kilometers from Athens, a prison envisioned to be a "disciplinary center" for political terrorists, like the founders of the neo-Nazi party "Golden Dawn" who were also serving there (Σουλιώτης, 2020). This was seen by Koufontinas's lawyers and supporters as an unjust and illegal punishment, evidence of a vindictive state policy against him, particularly because of his participation in 1989 in the execution of journalist Pavlos Bakoyannis, the father of Athens' Mayor Kostas Bakoyannis and husband of Dora Bakoyanni, ex-minister of foreign affairs and sister of Greek Prime Minister Kyriakos Mitsotakis.

Koufontinas went on hunger strike on 8 January 2021, demanding to be transferred back to Korydallos, as the law initially stipulated. He ended his hunger strike on the 14th of March,

[3] https://www.state.gov/foreign-terrorist-organizations/.

lasting 66 days[4]; his life reached imminent danger after having suffered kidney failure (Associated Press, 2021). The story gained so much traction, that a Spanish MEP filed a question to the European Commission about Koufontinas's situation[5], as the thought of an inmate dying of hunger strike in a European Union member state was daunting.

The situation led to extreme polarization within Greek society: while polls were showing that as much as 60% to 70% of Greeks were opposed to satisfying Koufontinas's demands, thousands of leftists and anarchists rallied daily to his support in Athens and other big Greek cities (Kitsantonis, 2021). Notwithstanding the pressure, the government did not back down; Koufontinas eventually stopped the hunger strike and received immediate treatment prior to being transferred to Domokos's detention center. It should also be noted that an appeal filed by Koufontinas's lawyer to Greece's Supreme Court was rejected[6].

A Content Moderation Barrage

Amidst this extremely tense situation, many Greeks started posting messages about the case as it can be seen in Figure 1 below. Greek Facebook users noticed a trend: numerous accounts were either being temporarily suspended or had their content removed after posting either about the protests or in support of Koufontinas demand (but not in favor of his terrorist acts).

Figure 1 - Content referring to Koufontinas on Facebook from 01/01/2021 to 30/04/2021 - Source: CrowdTangle

[4] https://www.efsyn.gr/ellada/dikaiomata/285561_me-dilosi-toy-o-dimitris-koyfontinas-stamata-tin-apergia-peinas.
[5] https://www.europarl.europa.eu/doceo/document/E-9-2021-000827_EN.html.
[6] https://www.iefimerida.gr/ellada/ste-aperripse-oristika-aitima-koyfontina-gia-domoko.

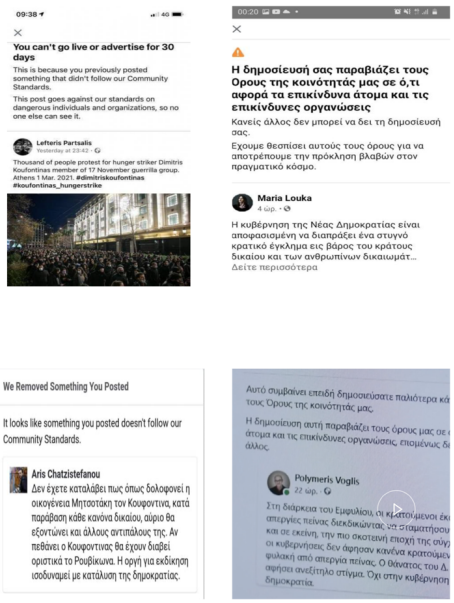

Figure 2 - Examples of accounts who have either had their content removed or accounts restricted

In this section, we will present empirical evidence concerning influential figures (i.e., journalists, photo-reporters, public personas and organization) that have had their account blocked or content taken down during the two months of Koufontinas's hunger strike (Table 1). The evidence was gathered through intensive web search and it is possible that there are cases of content moderation that did not receive similar attention as to the ones discussed here. Thus, it should be stressed that the selection of cases is not comprehensive but was rather based on the fact that the people or the organizations involved are public figures, as well as on the availability of information. Also, in many instances, the same content was posted on other platforms, namely Twitter, and no actions were taken against it.

Users	**Capacity**	**Policy penalty**	**Content**	**Policy Violation**	**Date of Action**	**Reinstated**
Aris Chatzistefanou	Journalist	Post removed	Opinion concerning the case and the government handling	Violation of Community Standards	23/02/2021	Unknown
Maria Louka	Journalist	Post removed	Opinion concerning the case and the government handling	Violation of Community Standards concerning Dangerous Individuals and Organizations	23/02/2021	Unknown
Konstantinos Poulis & The Press Project	Journalist	Warning of the page "being unpublished"; reduced distribution and other restrictions	Posted photos of demonstrations for Koufontinas's case	Recurring violations of Community Standards	27/02/2021	Unknown
Association of Greek Judges & Prosecutors	Association	Post removed	Official announcement concerning the case	Unknown	24/02/2021	No
Marios Lolos	Photo-reporter	Restriction of "going live or advertising" for 60 days	Photos they took of the demonstrations	Violation of Community Standards concerning Dangerous Individuals and Organizations	01/03/2021	Yes (02/03/2021)
Tatiana Mpolari	Photo-reporter	Post removed	Photos they took of the demonstrations	Violation of Community Standards	01/03/2021	Yes (02/03/2021)

				concerning Dangerous Individuals and Organizations		
Yannis Kemmos	Photo-reporter	Post removed	Photos they took of the demonstrations	Violation of Community Standards concerning Dangerous Individuals and Organizations	01/03/2021	Unknown
Lefteris Partsalis	Photo-reporter	Post removed	Photos they took of the demonstrations	Violation of Community Standards concerning Dangerous Individuals and Organizations	01/03/2021	Yes (02/03/2021)
Thanassis Kampagiannis	Lawyer/public figure	Many restrictions on his account	Multiple posts & re-shares concerning the case	Unknown	28/02/2021	Unknown
Polymeris Voglis	Scholar/public figure	Restriction of "going live or advertising" for 30 days	Opinion concerning the case and the government handling	Violation of Community Standards concerning Dangerous Individuals and Organizations	28/02/2021	Yes (01/03/2021)

Table 1 - Instances of content moderation actions related to Koufontinas's case

By looking at the instances mentioned above, we can infer the following:

i. The content that was moderated referred to the Koufontinas case.
ii. All posts were published between 23/02 and 02/03, a period included in the first peak of mentions of the term "Κουφοντίνας" (Koufontinas in Greek) in Facebook posts as shown in Figure 1; what is more, multiple demonstrations took place during that period. The second peak seen in Figure 1 is due to Koufontinas announcing the end of his hunger strike.
iii. Six out of ten posts were deemed to have violated Facebook's Community Standards in regard to its particular Dangerous Individuals and Organizations policy, while two others apparently violated Facebook's general Community Standards even though they referred to the same case; for the rest we could not find this information.
iv. Possibly, more than half of the actions were reinstated after user appeals.
v. Eight out of ten posts were made by left-leaning journalists, which could be a possible explanation of why they were targeted by Facebook's moderation system. However, the case of historian Polymeris Voglis and that of the Association of Greek Judges & Prosecutors is not in line with this hypothesis.

By analyzing the cases on Table 1, we have not identified any violations of Facebook's Community Standards. All of them were either reporting on demonstrations that were taking place at several cities in Greece, which are newsworthy events, or discussing the case's critical stakes without appraising Koufontinas's violent past as evidenced by our examination. This was also the opinion of a Facebook partner: Dimitris Alikakis, senior editor of "Ellinika Hoaxes," a certified Facebook fact-checker[7], said that "Facebook has made a mess of it all" and, referring to Mr. Voglis's case, that "it doesn't violate any policy" [8].

[7] The second one is Agence France-Presse. Its partnership was announced on 25 May 2021 (The LiFO Team, 2021)
[8] https://info-war.gr/anthropino-cheri-piso-apo-ti-logokrisia/.

Additionally, another element of inconsistency is that, while there were two cases found by moderators to be in violation of Facebook's general Community Standards, the other six were accused of violating the platform's Dangerous Individual and Organizations policies. This may indicate that different human decisions were applied to the same kind of content, something that has been empirically attested by research in other contexts of "manual" content moderation (Smyrnaios & Marty, 2020).

A Political Tantrum

Avgi, one of Greece's oldest leftist newspapers with strong ties to the former ruling party of Syriza, asked Facebook about the increased content takedowns on 26 February 2021. Facebook replied almost immediately through its Greek Public Relations partner, "Hill+Knowlton":

> *"Terrorists, violent extremist groups and hate groups have no place on Facebook and Instagram. In accordance with our Dangerous Individuals and Organizations Policy, we prohibit members of terrorist organizations, such as Koufontinas, from using our platforms, and we also block posts that endorse or support these individuals and their actions whenever we become aware of them.* ***Taking a neutral stance regarding terrorist groups and informing the media about their actions is not against our rules****"* [9] (translation and emphasis ours).

Facebook's policy on Dangerous Individuals and Organizations mentioned in their reply merits unpacking[10]. Like the content moderators' hierarchy, this policy is divided into three tiers as well: "Tier 1 focuses on entities that engage in serious offline harms [including] terrorist, hate, and criminal organizations. We remove praise, substantive support, and representation of Tier 1 entities." What is more, the platform heavily relies on the US government to define terrorists: "Tier 1 includes […] terrorist organizations, including entities and individuals designated by the United States government as Foreign Terrorist Organizations (FTOs)." In other words, "Silicon Valley tech companies rely on the US government to define 'terrorism,' and remove content from groups on the State Department's [FTO list]" (York, 2021, p. 126). Tier 2, then, focuses on violence against military and/or state actors that "do not generally target civilians," while Tier 3 "focuses on entities that may repeatedly engage in violations of our Hate Speech or Dangerous Organizations policies on-or-off the platform or demonstrate strong intent to engage in offline violence in the near future." Consequently, it seems that Facebook is following US policies without questioning, while promoting their agenda on a global scale. So, this policy has direct consequences in the structure of the political debate in third countries where the characterization of "terrorist" is contested (for example PKK in Turkey or Hezbollah in Lebanon).

Moreover, relevant content about the Koufontinas case that was posted from accounts and pages that were less left-leaning and/or militant was not removed, indicating the possibility of a political selective bias or targeted massive reporting campaigns (Smyrnaios & Papaevangelou, 2020). Facebook generally denies that mass reporting influences its moderation process; its Greek representatives repeated as much in their response to another request sent by the investigative journalists of *Reporters United*, claiming with emphasis that

[9] https://www.avgi.gr/social/380532_giati-afairoyntai-anartiseis-gia-ton-dimitri-koyfontina.
[10] https://transparency.fb.com/en-gb/policies/community-standards/dangerous-individuals-organizations/.

"even a single report is enough to partially block content, if it's confirmed that it violates [Facebook's] policies"[11]. More importantly though, in their response, Facebook vicariously admitted to having mistakenly enforced policy actions by stating that "although great progress is being made in our [content moderation] systems, they are still not perfect, and sometimes mistakes are made, **as in this case**. The content that was accidentally removed was reinstated when it was discovered" (Reporters United, 2021; translation and emphasis ours).

In their reply, the PR firm that acts on behalf of Facebook reaffirmed that the platform "allows people to neutrally discuss these individuals and share news reports about their actions." However, it failed to provide the public with more information as to what kind of mistakes were made and at which stage of the content moderation process in this particular case. It is interesting to note that the reply was published just one day after Facebook took action against many photo-reporters' content, causing the Union of Greek Photo-reporters to issue a reproachful announcement on the 2nd of March, calling out Facebook's for "censorship"[12]. The same strong language was used in an official letter of inquiry that the major left-wing opposition party in Greece, Syriza, sent on the 5th of March to the Prime Minister's office through the Hellenic Parliament concerning "censorship on social media"[13]. Also, Facebook did not provide an answer to *Reporters United*'s question about whether they had been contacted by the Greek government regarding this issue and whether Teleperformance was indeed conducting content moderation for Facebook in Greek language as it was reported in the press (Παπαδόπουλος, 2019). A Syriza MP also sent a letter to Teleperformance Greece asking for the company to take responsibility for the moderation errors that were reported[14]. He went even further by linking them to the fact that Teleperformance Greece was awarded with a public contract for providing call center services related to the Covid-19 pandemic and implied that censoring posts of left-wing journalists about Koufontinas could be a favor to the government in exchange for this contract.

Thus, the lack of clarity and transparency on behalf of Facebook leaves ample space for political speculation and, even, conspiracy theories. Furthermore, the double standards that were applied on similar content demonstrate further inconsistency and raise more suspicions of political bias. These suspicions are reinforced by the fact that a current Senior Communications Manager of Hill+Knowlton, the PR firm that represents Facebook in this affair, Mr. Leonidas Marcantonatos, is a member of the ruling party and holds the position of "Diaspora Affairs Lead for US and Canada" since May 2020 according to his LinkedIn page. This does not amount to evidence of governmental interference, but raises deontological eyebrows, especially since the company acts on behalf of a private conglomerate in Greece, creating thus the conditions for what scholars have called "regulatory capture" (Novak, 2013; Stigler, 1971).

Discussion

[11] https://www.reportersunited.gr/4153/ypothesi-koyfontina-to-facebook-tora-leei-oti-ekane-lathos/.
[12] https://www.avgi.gr/social/380862_logokrisia-kataggellei-i-enosi-fotoreporter.
[13] https://www.hellenicparliament.gr/UserFiles/c0d5184d-7550-4265-8e0b-078e1bc7375a/11588045.pdf.
[14] https://www.2020mag.gr/politics/1531-epistoli-vernardaki-stin-teleperformance-gia-facebook-logokrisia-se-sxolia-anartiseis-kai-dimosieyseis-pou-einai-epikritika-pros-tin-kyvernisi.

We have no concrete evidence to safely infer conclusions except for the empirical material presented in this chapter. If we were to speculate, we would argue that, first, Koufontinas is indeed blacklisted as a Dangerous Individual on Facebook; possibly because he and 17 November had been in the State Department's list until 2015. Second, we would argue that, given the traction that the case was receiving near the end of February, as can be seen in Figure 1, it is possible that people started reporting relevant content and accounts who disseminated it, triggering Tier 3 human moderators to review it. Third, these moderators saw that Koufontinas was a blacklisted terrorist and enforced policy without looking at the content's context. Fourth, as the suppressed accounts belonged to journalists and, thus, were quite vocal not only on Facebook but also on other media, the restrictions gathered a lot of attention and required the intervention of Tier 2 moderators. Last, Tier 2 moderators verified that no policy was violated and, to control damage, ordered the content to be restored (some of it was restored immediately or with a delay, but some of it was never restored).

Two questions remain: first, why were these accounts targeted when other news media or journalists were covering the story as well? And, second, what constitutes neutral discussions for Facebook if factual reporting can be taken down with such ease? A probable answer to the first question is that people, who did not want to see Koufontinas's demand granted or who disagreed with what was being posted, reported relevant content and accounts, either on their own or in coordination with others; whether there was an automated filter applied to content related to Koufontinas is uncertain but it is unlikely as that would have caused all relevant content to be restricted, as mentioned earlier. Furthermore, given the close interest of the Greek government in this case, a political intervention against specific content and/or accounts cannot be excluded (neither proved).

The second question is a bit trickier to answer because, theoretically, any piece of content that covers a controversial matter may be massively reported. So, it is possible that there is a loophole within the current content moderation process that requires adjustment: instead of relying solely on how news content is perceived by users to train machine learning models, there should be checks and balances to ensure that reporting and political discussions, even around controversial issues, should be left intact unless deemed harmful by properly trained human reviewers. This should also deter bad faith actors from gaming the report mechanism to their interest (Crawford & Gillespie, 2016).

Conclusion

To summarize, this chapter provided an overview of a case study of a content moderation debacle on Facebook concerning the news coverage and political discussion of a complicated case around a convicted terrorist's protest. In addition, we attempted to describe how Facebook polices content on its platform based on a three-tier system, along with a combination of human reviewers and algorithmic processes.

Our case study demonstrates that content moderation remains highly problematic for major social media platforms, despite their recent efforts to up their game. Anecdotally, Kate Klonick wrote in a Twitter thread that "for every 1.3 content moderation workers at FB there are 100,000 FB users"[15], highlighting the fact that while Facebook has invested the most

[15] https://twitter.com/klonick/status/1440363960851582981?s=21.

resources in content moderation among other Big Tech companies, the resources allocated are still incredibly low compared to what is needed.

We argue that this case study further demonstrates platforms' preference to proactively err on the side of more content removal (Douek, 2020) rather than letting controversial content on its services; this also further strengthens the argument that platforms are not neutral intermediaries (Gillespie, 2010) and are increasingly taking editorial decisions (Napoli & Caplan, 2017). Moreover, we posit that this is the path of least resistance for platforms, as it allows them to not tackle hard-pressing issues concerning freedom of expression and freedom of information. In other words, these companies have the necessary resources to invest in content moderation services that would align with their popularity and importance; the question is whether they have the will to do so. And, if not, whether public authorities have the political will to pass regulations that would force their hand and change the current content governance structure.

Last, through our investigation and relevant content analysis, we identified some issues that raise political questions about platforms' relationship with state actors, namely Facebook and the Greek government. In a time where tech lobbying has become the biggest lobby sector in the EU with a staggering annual record of €97 billion (Dr. Bank et al., 2021), alongside the ever-present phenomenon of revolving doors, where state actors take up crucial positions in private companies and vice versa (Stiglitz, 2014), cases like the one of Mr. Marcantonatos (Figure 3) do not help build trust in either public or private sector and can only worsen Greek citizens' trust in news outlets, which was already low according to the latest Reuter's Digital News Report, as "fewer than a third (30%) [of Greeks] trust the news" (Newman et al., 2020, p. 72). Deontological questions that are inherently connected to political integrity should not be left unanswered especially in the context of the Covid-19 crisis where many countries like Greece face threats on media freedom and pluralism.

References

Associated Press. (2021, March 5). *Greek militant on hunger strike suffers kidney failure*. PBS NewsHour. https://www.pbs.org/newshour/world/greek-militant-on-hunger-strike-suffers-kidney-failure

Boeder, P. (2005). "Habermas' heritage: The future of the public sphere in the network society. *First Monday*, *10*(9).

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

CrowdTangle Team. (2021). *CrowdTangle*. Facebook, Menlo Park, California, United States. https://www.crowdtangle.com

Douek, E. (2020). Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3679607

Dr. Bank, M., Duffy, F., Leyendecker, V., & Silva, M. (2021). *The lobby network: Big Tech's web of influence in the EU*. Corporate Europe Observatory and LobbyControl e.V.

https://corporateeurope.org/sites/default/files/2021-08/The%20lobby%20network%20-%20Big%20Tech%27s%20web%20of%20influence%20in%20the%20EU.pdf

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 205395172094323. https://doi.org/10.1177/2053951720943234

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., & Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, *5*(4), 529–538. https://doi.org/10.1038/s41562-021-01079-8

Hern, A. (2021, March 23). Decoding emojis and defining 'support': Facebook's rules for content revealed. *The Guardian*. https://www.theguardian.com/technology/2021/mar/23/decoding-emojis-and-defining-support-facebooks-rules-for-content-revealed

Horwitz, J. (2021, September 13). Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt. *Wall Street Journal*. https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353

Kassem, M., & Marwa, F. (2021, May 7). Sheikh Jarrah: Facebook and Twitter systematically silencing protests, deleting evidence. *Access Now*. https://www.accessnow.org/sheikh-jarrah-facebook-and-twitter-systematically-silencing-protests-deleting-evidence/

Kitsantonis, N. (2021, March 3). Protests and Vandalism Follow Hit Man's Hunger Strike. *The New York Times*. https://www.nytimes.com/2021/03/03/world/hunger-strike-protests-greece.html

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *HARVARD LAW REVIEW*, *131*, 73.

Napoli, P. M., & Caplan, R. (2017). Why media companies insist they're not media companies, why they're wrong, and why it matters.pdf. *First Monday*, *22*(5). https://doi.org/10.5210/fm.v22i5.7051

Newman, N., Richard, F., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Reuters Institute Digital News Report 2020* (p. 112). Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Newton, C. (2021, September 23). *Why these Facebook research scandals are different*. The Verge. https://www.theverge.com/2021/9/23/22688976/facebook-research-scandals

Novak, W. J. (2013). A Revisionist History of Regulatory Capture. In D. Carpenter & D. A. Moss (Eds.), *Preventing Regulatory Capture* (pp. 25–48). Cambridge University Press. https://doi.org/10.1017/CBO9781139565875.004

Reporters United. (2021, March 3). *Υπόθεση Κουφοντίνα: Το Facebook τώρα λέει ότι «έκανε λάθος»*. Reporters United. https://www.reportersunited.gr/4153/ypothesi-koyfontina-to-facebook-tora-leei-oti-ekane-lathos/

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Singh, S. (2019). *An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content* (p. 42). New America: Open Technology Institute. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

Smyrnaios, N., & Marty, E.(2020). Occupation:"Net Cleaner"—The Socio-economic Issues of Comment Moderation on French News Websites. In Ballarini, L. (Eds.), *The Independence of the News Media Francophone Research on Media, Economics and Politics* (pp. 103-131). Palgrave Macmillan.

Smyrnaios, N., & Papaevangelou, C. (2020, December 10). *Le signalement sur les réseaux sociaux, un moyen de modération mais aussi de censure*. La Revue Des Médias. https://larevuedesmedias.ina.fr/signalement-reseaux-sociaux-moderation-censure

Statista. (2021). *Most used social media 2021*. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Stigler, G. J. (1971). The Theory of Economic Regulation. *The Bell Journal of Economics and Management Science*, *2*(1), 3. https://doi.org/10.2307/3003160

Suzor, N. (2020). Understanding content moderation systems: New methods to understand internet governance at scale, over time, and across platforms. In R. Whalen, *Computational Legal Studies* (pp. 166–189). Edward Elgar Publishing. https://doi.org/10.4337/9781788977456.00013

The LiFO Team. (2021, May 25). *Facebook: Μόλις ανακοίνωσε συνεργασία με AFP για αποτελεσματικότερο έλεγχο των fake news στην Ελλάδα*. LiFO. https://www.lifo.gr/now/tech-science/facebook-molis-anakoinose-synergasia-me-afp-gia-apotelesmatikotero-elegho-ton-fake

The Manifold. (2020, August 6). The Covid-19 crisis highlights Greece's media problem. *International Press Institute*. https://ipi.media/the-covid-19-crisis-highlights-greeces-media-problem/

York, J. (2021). *Silicon values: The future of free speech under surveillance capitalism*. Verso.

Zuiderveen Borgesius, F. J. (2020). Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights*, 1–22. https://doi.org/10.1080/13642987.2020.1743976

Μάνδρου, Ι. (2019). *Ολες οι αλλαγές στον Ποινικό Κώδικα | Η ΚΑΘΗΜΕΡΙΝΗ*. ΚΑΘΗΜΕΡΙΝΗ. https://www.kathimerini.gr/politics/1049697/oles-oi-allages-ston-poiniko-kodika/

Παπαδόπουλος, Γ. (2019). *Στα άδυτα του Facebook στο Μοσχάτο | Η ΚΑΘΗΜΕΡΙΝΗ*. https://www.kathimerini.gr/investigations/1047933/sta-adyta-toy-facebook-sto-moschato/

Σουλιώτης, Γ. (2020). *Μεταγωγή Κουφοντίνα μετά την ψήφιση του νέου νόμου | Η ΚΑΘΗΜΕΡΙΝΗ*. ΚΑΘΗΜΕΡΙΝΗ. https://www.kathimerini.gr/society/561203578/metagogi-koyfontina-meta-tin-psifisi-toy-neoy-nomoy/