RIVISTA INTERNAZIONALE DI FILOSOFIA E PSICOLOGIA
DOI: 10.4453/rifp.2020.0021
ISSN 2039-4667; E-ISSN 2239-2629
Vol. 11 (2020), n. 3, pp. 307-326

RICERCHE

Commercial Content Moderation: An opaque maze for freedom of expression and customers' opinions

Paolo Petricca [(α)]

Ricevuto: 9 dicembre 2019; accettato: 18 novembre 2020

Abstract The present work analyses Content Moderation, focusing on ethical concerns and cognitive effects. Starting from a general description and history of the moderation process, it stresses some ethical problems: quality of moderation, transparency, and the working conditions of human moderators. Using some of Facebook leaked slides offering examples of moderation, we define some controversial rules and principles for Commercial Content Moderation. These examples highlight a general lack of coherency and transparency, which has the potential to affect users' cognitive attitudes, their perception of reality, and their freedom of speech. Such effects are studied in comparison to other well-known online cognitive phenomena (bubbles and echo chambers) and in relation to the most recent dedicated legislation in EU countries. The current Content Moderation scheme leaves users at risk of specific cognitive distortions, highlighting the urgent need for greater transparency throughout the moderation process and better working conditions for moderators.
KEYWORDS: Content Moderation; Freedom of Speech; Epistemic Bubble; Technology Ethics

Riassunto *Commercial Content Moderation: un oscuro labirinto per la libertà d'espressione e le opinioni degli utenti* - Il presente lavoro offre un'analisi approfondita della *moderazione di contenuti*, concentrandosi sulle problematiche etiche e sugli effetti cognitivi. A partire da una introduzione ai concetti chiave e alla storia della moderazione di contenuti online, si concentra su alcuni problemi etici primari: la qualità della moderazione, la sua trasparenza e le condizioni di lavoro dei lavoratori. Utilizzando dei documenti formativi interni di Facebook, pubblicati da un quotidiano e dagli esempi in essi contenuti, definiremo le controverse regole e i principi della *Commercial Content Moderation*. Ne emerge una generale mancanza di coerenza organizzativa e di trasparenza nel processo, che mostra potenziali effetti dannosi sulle attitudini degli utenti, sulla loro percezione della realtà e la loro libertà di parola. Tali effetti saranno studiati in confronto con i più conosciuti effetti cognitivi del mondo dei social (*epistemic bubble, filter bubble, echo chamber*) ed in relazione alle più recenti disposizioni di leggi europee in merito. Gli schemi attualmente in uso producono degli specifici effetti di distorsione cognitiva e conseguentemente mostrano l'importanza di una maggiore attenzione alla trasparenza del processo e alle condizioni di lavoro dei moderatori.
PAROLE CHIAVE: Moderazione di contenuti; Libertà di parola; Bolla epistemica; Etica della tecnologia

[(α)]Dipartimento di Lingue, Letterature e Culture Moderne, Università degli Studi di Chieti "G. D'Annunzio", viale Pindaro, 42 - 65127 Pescara (I)

E-mail: paolo.petricca@unich.it

1 Introduction

THIS WORK WILL ASSESS COMMERCIAL Content Moderation (CCM) as a crucial component in a safe online environment, which involves complex and opaque dynamics with the potential to influence users' freedom of expression and communication practices. The enquiry starts with a presentation of moderation practice, including key definitions, a brief history, and outlining its main characteristics. We then provide an overview of the three main ethical issues involved in Content Moderation: censorship, transparency, and the critical issues affecting moderators.

The analysis will then focus on moderation rules and examples, based on official training slides leaked by Facebook employees. This will highlight the main potential problems with Content Moderation: freedom of speech is limited by legal dynamics and overall lack of competence among moderators. The widespread nature of these risks will be shown through the analysis of several important instances of content removal, national legal interventions, and examples of moderation that fail to clarify rules.

2 Commercial Content Moderation

In this section, we provide some basic definitions, discuss the main characteristics of Content Moderation, its historical development, and point out some important features. The topic is very complex so this introduction will not seek to provide a complete presentation, but necessarily remain incomplete and focused on our research objectives.

2.1 Content Moderation

In order to provide a well-balanced and useful presentation of Content Moderation, we start with some basic definitions. A *Content Platform* (CP) is any online site that hosts shared content and social interactions among users;[1] that content needs to be produced by spontaneous activity within the community, which is not compensated or commissioned by the platform itself. The platform must be built on an infrastructure that processes data and provides customer service. Within this comprehensive definition we can include social networks (Facebook, Instagram, Twitter, etc.), search engines (Google, Yahoo, etc.), and content collectors (Wikipedia, Reddit, 4chan, etc.).

Content Moderation (CM) activities are practices aimed at rule-based governance of content. Rules allow moderators to perform three fundamental functions: preserve a certain standard of the general quality in terms of content truthfulness; avoid potential personal and social harm generated by content, either by removing or compelling the user to modify such content; punish authors of harmful content either by proposing internal disciplinary actions or by reporting them to the proper legal authorities.

Commercial Content Moderation (CCM), as proposed by Sarah T. Roberts,[2] includes the complex set of moderation practices exercised within a *platform* (generally, hosted by a major tech company), following its private rules and for its private objectives, by non-volunteer workers trained for this job.

2.2 A brief history of CM

For much of the last three decades, there was only mild interest in CM, and this task was fulfilled only by members of internet forums and communities as well as a few professional insiders. Online communities have been created since the very beginning of Internet history.[3] In such forums, a moderator was usually member, acting as first among equals, who followed just a few rules that were determined by the whole community and aimed to preserve the high quality of interactions between users and exchanged contents.[4] Moderators' activities were largely thought to fulfill the requirements contained in the famous Request for Comments (RFC) 1855, published in 1995, better known as Netiquette protocols.[5]

Over the years, however, many new issues have emerged: users have started to share more complex contents, not just in professional or thematic communities but also on general sites, thereby altering the objectives of CM; community membership can reach hundreds of thousands; users share nearly every kind of file. In the late 1990s, the main issues faced in moderation involved copyright, since many proprietary files were easily shared without any control over peer-to-peer protocols (P2P). Indeed, the major entertainment industries deployed their financial power in order to contain these acts of piracy.[6] During the 1990s and in the first decade of the 21st century, the number of Internet users grew to millions and Internet usage spread throughout the world, multiplying the ways and purposes for which online activities were used. With this growth, many political and social phenomena have increasingly spread via internet communities and sites: political ideas, sexuality, racism, faith – nearly every human activity has an online presence. This acceleration followed an exponential trend, and moderation started to become more widespread and professionally shaped. In 2007, sales of a new device called the smartphone surpassed those of other traditional internet devices;[7] in 2014, web searches from smartphones became the most common use of the Internet in the world. This further technological expansion increased the worldwide number of Internet users to billions and permanently changed the dynamics of the Infosphere[8].

The latest stage of this development, strictly related to CM, is the capillary-like diffusion and maintenance of social networks (SN). At present, more than 60% of populations in the world's industrialized countries use at least one social network.[9] The biggest social network, Facebook (FB), reports 2,603 billion active users per month.[10] With the rise of SNs, the whole vision of the Internet has changed: concepts like privacy, truth, trust, freedom, and even democracy have changed their shape and influenced public perceptions, raising many ethical issues. Overall attention to SNs has grown significantly over these years, because the sheer mass of worldwide users in these communities has magnified the potential effects of phenomena like fake news, trolling, profiling, and privacy infringements.[11]

The apex of all these public concerns has been reached with the Facebook-Cambridge Analytica Scandal in March 2018, had an enormous impact on public opinion as people witnessed the potential influence of a private company over the U.S. 2016 presidential election and other important elections. Focus on this scandal came to a head after two years of investigations by newspapers and several impactful reports.[12]

A similar dynamic occurred with respect to CCM, which had received only niche attention for several years but was exposed to increased public debate by several influential news reports (The Guardian, Gawker, The Verge and the New York Times), following major leaks of documents from FB's internal moderators. The anonymity of the inside sources and the secrecy which covers the moderation processes of all the major CPs contributed to keeping public attention to this scandal at low levels. Nevertheless, over the last five years, the number of books, news reports, official documents, and academic papers on this topic has constantly increased, shedding light on the many ethical problems and social consequences associated with CMM and its failures.

As documented in leaked documents, workers' testimonies, and extended studies, the conventional practice of CMM leads to a wide range of technical, economic, psychological, legal, ethical and cognitive problems. The present work is devoted to highlighting such issues and discussing potential methods to tackle them. We suggest a systematic revision of CMM best-practices and major involvement on the part of governments and public agencies, who should define the social responsibilities of CPs, and their duties with respect to their moderators and customers.

2.3 Characteristic

CCM has several technical problems. The first involves the scale of operation: FB, for instance, has two billion users who are responsible for billions of posts per day in more than a hundred languages. FB's CCM has to manage more than 10 billion posts per week and they aim to do this job with a maximum error rate of 1%, moderating all user-reported content within 24 hours.[13] FB, YouTube, and Google are, so far, the only platforms operating on such a large scale.[14] Users are content creators and their activity is strictly necessary to the existence of these CPs. With such high user numbers, a complete and efficient moderation process based solely on human input becomes nearly impossible.

A.I. can help address this problem of scale and make human work on moderation considerably easier and safer, reducing the levels of human exposure to harmful contents. Ideally, FB's goal for moderation is to have a giant set of rules and examples, covering every conceivable grey area, so that potentially dangerous content can be recognized automatically. This intent is technically difficult to achieve and leads to some paradoxical consequences. For instance, the aim of removing revenge porn content from the platform generated the following strange proposal: FB invited their Australian users[15] to privately upload their nude pictures to the platform so that their algorithms could learn to easily recognize if someone else had tried to post nude photos; user reaction was, predictably, not enthusiastic.

This case leads us to another, closely related, debate: A.I. or Human moderation? If we consider the data on the scale of the endeavor, the answer becomes obvious. FB, in 2020, can count on 15,000 human moderators based in the U.S.A. and many more worldwide;[16] the number has been constantly increasing. But this is insufficient to handle billions of posts: A.I. intervention is inevitable. The challenge is how to properly train AI algorithms – especially, considering that even human moderation remains very complex and nuanced. We can say that the presence of Machine Learning in CCM is necessary, but we must always remind ourselves that A.I. contributions are data-driven and can be tricked if malicious users know the moderation algorithms.

The first stage of CM is filtering: the software scans tons of daily posts and generates reports for human moderators. At the same time, any user can *flag* content as inappropriate and bring it to the attention of the moderation process. The role of such flagging is essential in moderation, but it can also have dangerous social and cognitive impact, because when massive human flagging operations are performed for ideological purposes, the CPs are not robust enough to regulate potential outcomes. In cases when, for example, a political party encourage its followers to flag content posted by the opposite party as inappropriate *en masse*, moderators tend to remove the content, even if it is not violating any rule, and only later perform a serious analysis on the content. This process could require hours, or even days, and when the content is (hopefully) restored, the political debate may already be focused on other events. As far as we know from documents and literature, there is no priority protocol for such cases.

At a second stage, all flagged materials that have been sent to moderators are outsourced to third parties, who can choose from three options: *ignore* the content, because it does not violate any rule; *delete* it, if it resembles cases they were trained to recognize; or finally *escalate* the content to a group of workers higher in the moderation hierarchy. If a decision is difficult, escalation may be performed recursively and forwarded to internal teams, senior teams, and high-level teams. We don't know much about the internal hierarchy of teams in CPs, but we know (directly from the platform) that FB has one team that specifically responds to content moderation crises. Other FB teams write moderation software tools, try to ensure ac-

curacy and consistency across the globe, and attempt to coordinate all of the other teams so they work together well.[17]

The highest-level team in FB, to whom content can be escalated is called Risk and Response. It works with policy and communication teams to make tough calls. Some journalists from Vice participated in an internal meeting where special cases, reported by the Risk and Response team, were debated and updated. This process is long and highly articulated: «Teams from 11 offices around the world tune in via video chat or crowd into the room. In the meetings, one specific "working group" made up of some of Facebook's policy experts will present a "heads up," which is a proposed policy change. These specific rules are workshopped over the course of weeks or months, and Facebook usually consults outside groups – non-profits, academics, and advocacy groups – before deciding on a final policy, but Facebook's employees are the ones who write it».[18]

The moderation process is the result of an articulated patchwork, which comprises many variables (*banning policies*, the use of *blocklists*, etc.) and possible outcomes, from the perspectives of the platform and the users. But we must keep in mind that, we only know some details about this process, thanks to leaked internal documents; the overview given by company executives contains few details and no explicit references to partners, real cases, aims, or leading principles. Opacity and secrecy make inquiry difficult and increase the risk of ethical issues.

2.4 Secrecy

As we can easily sense, CCM is a very complex phenomenon which features many problems of different types. This complexity is further increased by a serious lack of official data, documents, and testimonies on CCM; this scarcity has several causes. First of all, we must consider that, for many years, this practice was not imposed on CPs by law; the platforms have had their own interests in maintaining a moderated environment for their communities. This means there is no reliable way to assess if these interests are commensurate with their concrete social responsibility to control and influence people expressions and interactions. The similarity of this practice to censorship cannot be overestimated.

Scholars and journalists around the world are eager to know the details of CCM, but this wish is in continuous opposition to the inner nature and aims of CPs, which are for-profit private companies. Their main goal is to increase their profits and, within this perspective, they prefer to maintain a certain level of secrecy. This secrecy is protected by the use of general Community Guidelines and the use of non-disclosure clauses when hiring moderators.

Until leaked documents led to several relatively recent scandals, CCM was commonly considered to be an internal process. In 2017, several FB training slides were leaked,[19] putting moderation processes under a different light: FB was forced to talk about an issue that, until that moment, had been considered strictly confidential. The Cambridge Analytica scandal had reached fever pitch at that time and these disclosures attracted much attention to FB, leading to growing distrust of their policies and procedures. Their line of defense was very similar in these two different cases: they "naively" denied any knowledge or participation in all those activities that were not directly attributable to their company; they described and confirmed their support for freedom of speech and the well-being of their users, claiming this was guiding principle of the company.

In order to fulfill requests for clarity on moderation rules, on April 24, 2018, FB released their Community Standards,[20] 25 pages of general principles regulating the main concerns of moderation: Violence and Criminal Behaviour, Safety, Objectionable Content, Integrity and Authenticity, Respect of Intellectual Properties. But if the users expected to receive clear distinctions and some

key examples, like those in the leaked slides, they were surely disappointed. FB followed the path laid out by other major CPs (some of them owned by FB itself): Community Standards are general expectations about appropriateness, harassment, offensiveness, and prohibitions; explanations are provided, immediately following each general principle, by a list of things that are allowed and forbidden, while justifications are often considered unnecessary. The level of detail provided does not allow for specific analyses of the moderation process and does not refer in any way to the actual procedures carried out by human moderators or by algorithms.

The other, more powerful tool of secrecy is the non-disclosure agreement signed by hired moderators; the clause is valid for both internal CP workers and workers in companies to which moderation is outsourced. Several journalists have interviewed anonymous employees from CPs about their working conditions and moderation procedures, but the most authoritative and complete study on this topic is the book *Behind the Screen* by S.T. Roberts.[21] In this study, the moderators are shown to be workers with short contracts, committed to working an eight-hour day, who watch and read content that is [suspected of being] in violation of rules and are given highly restricted timeframes for taking their decisions. Many of the external companies hire workers from Latin America, Africa, and Southeast Asia (especially the Philippines). Many of the moderators work in call-centres or on micro-labour platforms (e.g., Amazon Mechanical Turk), without psychological assistance despite their constant exposure to potentially traumatic contents; in many cases, they work alone, in their homes, without any direct or easy contact with colleagues or superiors. A young Moroccan employed by one of FB's third companies (oDesk) revealed some details of his employment to A. Chen:[22] he was paid just $1 per hour, his content-moderation team used a web-based tool to view a stream of pictures, videos, and wall posts that had previously been reported by users. Many workers interviewed by Chen were afraid to talk because they had signed non-disclosure agreements; they suspected the interviews were just another test of loyalty commissioned by the company.

FB and CPs in general might be embarrassed if information on their moderation process, working conditions, and other details were to be made public. This is one, but not the only reason, for their maintaining a certain opacity about CCM practices. Other reasons for this choice will be discussed in Section 3, but it is appropriate to stress here one important consequence of such secrecy. CCM is a complex and delicate aspect of social networks and individual and collective online life. Opaque practices leave many of our hypotheses without a strong empirical confirmation: any study or ethical theory about CMM can rely only on general guidelines, interviews, or leaked documents (often revealing content that is suddenly replaced by the company). The policy of secrecy has often frustrated journalistic enquiries and academic research on this topic, allowing these companies to operate far from prying eyes.

3 Ethical issues

3.1 Content Moderation as censorship

Many people are concerned about the idea that CCM is de facto censorship exercised by CPs which affects user content; indeed, this is often the basis for public interest in the topic. In a general sense, censorship is carried out by an authority and it is easy to conceive of this term being used in relation to a public institution. But if we consider the CP as a private subject exercising active control over something that is legitimately within their control, a small minority of people would still define this as censorship. Probably, we would employ different terms, such as filtering, controlling, quality checking, or protecting interests. This situation raises an ethical issue: is CMM legitimate? Yes, absolutely.

When we post something on FB and share it with the public, according to the terms of

service, we cede our rights to the platform and all its users. Our content becomes a kind of open-access material, computationally managed by FB's license. If I publish a photo, any allowed viewer is a potential user and even if I were to eventually remove that photo myself, it would continue to be accessible to users in accordance with FB's License: Facebook's IP license still applies if «your content has been shared with others, and they have not deleted it».[23] Legal issues related to CPs are a fascinating topic, but one that lies beyond our present concerns. Returning to our theme, we – as users who have signed the terms of service – cannot legitimately contest FB authority by claiming it is a form of censorship.

Nonetheless, as citizens, we can and should be worried about the social outcomes of CCM because it affects the balance of various social dimensions. First of all, CPs can become so extensive that they effectively represent a sort of second society that is specular but not parallel to our real-world society. In this second society, FB and a few other CPs act like giant empires operating in their own interests and against their competitors. Actions undertaken in a CP's interest can have huge impacts on the real world. In March 2018, U.N. investigators reported that FB had direct responsibility for the Myanmar crisis:[24] FB had decided not to moderate some content exhibiting hate speech and this influenced the course of a major political crisis, incurring many casualties. In Section 6, a deeper argumentation addressing this near-censorship by CCM will be put forward, but there is a particular phenomenon related to CCM worth mentioning.

The two largest CPs, FB and YouTube, are enormously influential editorial content managers: even if they do not produce content themselves, they are respectively the biggest news media and broadcasting platforms in the world. Their CCM rules influence their editorial policy, in areas where local laws on internet censorship are either present or absent. These editorial powers have come to define new standards about what is admissible or not, filtering a huge portion of all news and videos on the Internet. Not only does every user willing to put his content on the CP have to follow these rules, so does every professional newspaper or broadcaster who intends to directly or indirectly share their content on the platform. CPs are like gargantuan service providers, so big and powerful that they can impose their rules on their clients. This may prove very problematic, because the clients (active users) constitute 31,3% of the human population.

3.2 Transparency in CCM

Since the first scandals damaged FB's public image, the platform's founder Mark Zuckerberg has made more personal appearances with the aim of defending the company's behaviour and restoring its credibility. One of his communication strategies has been to personally write a number of public letters to users and employees. In one of the first letters, sent out on February 2016,[25] he wrote about CM and its aims and methods in the future.

Aware of the widespread public perception of CM as censorship, the FB CEO is preconizing a future in which people of different countries explain the types of content that they want to see, in a participative way, and A.I. ensures that they are shown only the contents approved by their own sensibilities. He describes this as a «large-scale democratic process to determine standards with A.I. to enforce them» and added: «For those who don't make a decision, the default will be whatever the majority of people in your region selected, like a referendum». This democratic view of future CCM ends with this declaration of intent: «We are committed to always doing better, even if that involves building a worldwide voting system to give you more voice and control». The company knows that many western users might disdain the CCM model, based on its opaque processes and justifications, unclear rules, and the lack of any direct participation on the part of users. The actual CCM process is

far from Zuckerberg's vision and nearer to users' fears.

Our ethical concerns should not be considered only under the users' perspective and we should ask ourselves if transparency is an absolute value. Globally, freedom of speech and the freedom to share internet content are governed by national laws that differ from country to country; in this many-faceted situation, a CP needs to be conformed to each country's legal regime, in order to avoid issues with national authorities. Of course, in countries where transparency is not considered a virtue, a CP will not even consider publishing the rules and processes used for moderation. On the other hand, in a hypothetically completely transparent version of CM, CPs would activate CCMs in fewer cases but would be exposed to legal action and potential debates reflecting public opinion on their procedures and principles. This could slow down the moderation process and make it more expensive. However, there would be another even greater risk for CPs: moderation, whatever the principles and aims, would be at risk because, by making their rules public, they would give malicious users information that would allow them to cheat by elaborating precise and effective strategies to bypass these specific rules.

In order to avoid the potential problems of this "transparent version" in the real world, FB tries to keep the CM process as opaque as they can. We cannot speculate on Zuckerberg' sincerity and even less evaluate the company's appraisal of transparency. A more helpful attitude would be to consider the role of transparency at certain specific stages of the CM process; this will be outlined in Section 4.

3.3 Working conditions and skills required of human moderators

Above, we noted that one of the earliest CCM issues to draw attention was the working conditions of human moderators. It is difficult to explain why, but this third major ethical issue has since garnered less attention than the others. Of course, the threat to the mental and physical health of CCM workers is very real. This job can be considered arduous and exposes workers to considerable psychological trauma. Many symptoms of PTSD and other major mental disorders have been recorded and published in anonymous interviews by The Verge and The Guardian, by the book *Behind the screen*[26] and by reports, and include other CPs[27] besides FB. Many workers are young students trying to raise money for their university education and young single mothers from poor countries. This ethical issue is not limited to the plausible damage produced by the specific nature of this job, but also includes more general working conditions: very long working session, extremely low wages, and, in some cases, the isolation of teleworkers.

All of these adverse conditions are worrying in terms of the effects they can produce on workers and are problematic in terms of moderation quality. There are testimonies from workers involved in classifying and tagging suspect and dangerous posts[28] on the pressure they endure to make decisions quickly and to be productive. It seems that the web tools used by moderators have been, and possibly still are, complex and intricate, wasting time and lowering their productivity. Some workers report estimated decision times of around 8 seconds per content, others say 20 seconds, with the maximum estimate being around 30 seconds. We do not know if there is any quality control or supervision moderation of or, if so, how it works; but, since these working conditions are the result of policies aimed at reducing expenses so as to maximize CP incomes, we can legitimately suspect that there is no quality control. Probably, the few cases where the outsourced work is checked occur when moderation leads to the maximum penalties: banning or appealing to public authorities. By contrast, we should not expect forms of supervision for all the cases in which the content is simply ignored or removed.

The last aspect to consider is that, although testimonies from employees at different CPs sometimes differ making generalisation difficult, procedures for staff selection do not appear to be uniform. This is easy to infer from the fact that most workers are hired by external companies located in different countries and continents. Moreover, the low salaries lead us to think that this selection processed is not aimed at – nor will it attract – the most skilled workers on the market; even if there are areas in the world where labour costs are very low, and a company could easily hire a graduate employee with certified multilingual skills for a low wage, it is difficult to imagine that this worker would accept a payment of $1-4 per hour, enduring this particularly stressful type of work for years.

All of these ethical issues pose serious threats to workers' health and the quality of moderation.

4 Analysis: Leaked guidelines and legal standards

4.1 General policy

FB documents leaked to the British newspaper, *The Guardian*,[29] reveal many guidelines, rules of thumb, and examples of CMM in several content areas. Typically, they contain definitions about dangerous and offensive content, set specific limits (on sexual and other sensitive content) that should not be exceeded, and cutting-edge examples, useful for defining a large number of possibly dangerous posts.

For instance, the revealed slides showing animal violence try to define the limit of what is considered disturbing or not for users. The platform allows posts with such representations if they aim to increase awareness and keep users informed; if a post includes images, video, or text that promote the sadistic, ruthless, or unjustified use of violence, it will instead be considered a potential offence to users and be removed. This kind of distinction, whether it is detailed or stipulated by general or indirect principles, is the most common kind of content found in FB's slides for moderators. The slides about animal violence,[30] graphic violence, non-sexual child abuse, credible threats of violence, and «sextortion» or «revenge porn», all define these kinds of limits, based on reasonable protocols to ensure public security and respect common decency and sensibilities.

The slides were clearly intended as guidelines to be presented at seminars for moderators or as updates of more general principles. The examples do not always clearly present the principles behind the rule; on several occasions, the authors admit to having difficulty establishing clear guidelines and setting limits. Some important slides referring to critical CCM issues are examined below to highlight specific weaknesses in the moderation process, in particular, in relation to the FB platform.

4.2 Geo-Blocking and Holocaust denial

In the slides dealing with Holocaust denial,[31] we see further examples that reveal FB's ideas on freedom of speech and how to handle user sense of decency. FB decided to let individuals and groups make comments which support or advocate Holocaust denial, except in countries where this activity is actually considered illegal. More precisely, there are 14 countries that consider this practice illegal, but just 4 of them (Germany, France, Austria and Israel) officially requested online moderation of this content. This situation leads to Geo-Blocking: these contents will not be deleted from the CP, but just made inaccessible to users connecting from those 4 countries.

The leaked documents say such practices are necessary to avoid the risk that FB is banned in countries where certain contents posted on the platform are considered illegal; the countries where FB considers itself most at legal risk are Germany, Turkey, Pakistan, India, and Russia.

The most important case of Geo-Blocking took place in Pakistan in 2010,[32] as a result of

a snarky challenge called Everybody Draw Mohammed Day. This project was initiated by a group of U.S. users long before the Charlie Hebdo attack in 2015. The group wanted to challenge the Muslim prohibition on depicting the Prophet by gathering a critical mass of users who would draw and post an image of Mohammed. This provocation was intended to involve so many people that no one could effectively be threatened by terrorist retaliation. It attracted considerable attention from both western users and journalists who were willing to participate – of course, it also attracted attention from authorities in Muslim states. The group grew to 100,000 users by the time it was announced that the day of action would be the 20 May 2010. At the time of this announcement, the national High Court ordered the Pakistan Telecommunications Authority to block access to all Facebook sites, as well as YouTube, and the pertinent pages of Wikipedia and Flickr for Pakistani users.

At that point, FB found itself at a crossroad: to stand up for freedom of speech and expression, indirectly endorsing this provocative act and face historic legal action, or to indulge the High Court's request by removing the group from the platform.[33] In cases like these, we can see how difficult CCM can be for a CP: their activities can clash with public interests, laws, and morality. CPs face difficult challenges and sometimes are not prepared or structured to handle these ethical issues. In this case, the company decided to geo-block all contents from this group for all Pakistani IP. This answer was not a kind of egg of Columbus nor a Gordian knot, simply an attempt to manage a complex situation, satisfying both the contenders. Notably, however, FB refused to take a decision which would have defined the fine line between freedom of speech and social morality, preferring to consider this issue outside its area of responsibility. This is a key example showing how CPs are often asked to provide rules for ethical issues, although they are not designed to fulfill this role. Honestly, considering FB is a CP, we cannot blame it for making the choice it made. But if we think about the number of users involved and the potential outcomes of FB's approach to moderation, we can easily see how the importance of CCM activities can exceed the capacity of a simple CP.

A last interesting example showing how a CP can use geographical content moderation is in Google Maps. If a Russian user of the world-renowned mapping application looks for Crimea, they will find the region belongs to the Russian Federation; all other users will see it is part of Ukraine. The border is depicted as continuous in Russia but as a dashed line in other countries.[34] This is a clear example of how content management has to deal with pressing political interests and important disputes in the real world. All these cases show that CCM can have major effects on how users perceive reality, on their effective freedom of speech, and on public order.

4.3 Nudes and sexuality

FB's leaked policies on nudity[35] reveal a strange patchwork of rules of thumb. A very important case of unfortunate content moderation concerns the famous war photograph known as "Napalm Girl". This picture earned its author, Nick Ut the 1972 Pulitzer Prize; it is one of the most famous pictures from the Vietnam War. In September 2016, the Norwegian journalist, Tom Egeland, included it in an article reflecting on photographs that changed the history of warfare; his article was shared on FB, but the combination of suffering and underage nudity depicted in the photograph led the moderators to remove his post. After provocatively reposting the image, Egeland was suspended twice; many sympathetic Norwegian citizens and the Prime Minister herself, posted the same picture only to see if their posts were also removed, as they really were. This brought the case to a national audience, and a full-size version of the photograph appeared on the front page of the national newspaper Aftenposten on 8 September 2016, with a very

disapproving letter on the misapplication of moderation. As reported in Reuters,[36] the photo «had previously been used in training sessions as an example of a post that should be removed [...]. Trainers told content-monitoring staffers that the photo violated Facebook policy, despite its historical significance, because it depicted a naked child, in distress, photographed without her consent». This application of these rules was clearly not context sensitive.

In fact, the leaked slides on the Holocaust specify that nudity related to historic situations within the Holocaust (e.g., nudity in concentration camps) was allowed as a newly inserted exception to the policy on nudity. Unfortunately, due to the secrecy of the actual official guidelines for moderators, we are unable to analyze FB's full policy with respect to special cases of nudity. Nevertheless, these revealed cases confirm the lack of a coherent policy and the fact that the moderation system has been unable to distinguish contextualized usage, the purpose, or the value of nude pictures as historical documents.

In a similar way, imprecise and controvertible semiotic rules are applied more generally to nude images on FB. In the leaked slides on sexual activity,[37] we see an explicit group sex scene, with actors post-edited using pixelated graphics from a famous videogame and a male sexual organ depicted as a big mushroom (always inspired by the game), is allowed because there is no formal nudity even if the pornographic intent is obvious.[38]

The same rulebook produces the systematic removal of pictures from breastfeeding mothers where there is no sexually-oriented nudity and even if the photo is posted in a closed pro-breastfeeding group. One of the basic FB rules is that pictures cannot show any nude sexually-sensitive part of the body, but it can graphically represent those parts covering some detail and refer to them in any sexual attitude, without any problems. Over the last few years, the most critical and ironic users have inferred that FB is obsessed with prohibiting female nipples: if you are a breastfeeding mother you should avoid any square millimeter of nude nipple in your photos, but you can post a photo of female breast groping even in obviously sexual poses that show the whole naked bosom, provided that there is no nipple in sight. This obsession has also inspired some internet creativity, as some users have tried to ironically bypass such rules, suggesting the cunning substitution of male nipples for female ones.[39]

Even when nudity is represented in art, FB presents its moderators with questionable guidelines and examples. The main guidelines, as far as we know from the leaked documents, allow for nudity in handmade art but prohibit nudity in digital art; the explaining principle is: «We drew this line so that we could remove a lot of very sexual digital nudity but it also covers an increasing amount of non-sexual digitally made art».[40] In this case, and it is not the only one, the authors of the moderation policy are aware of limits to their rules; their attitude is to provide a general rule for moderators, making them aware of inherent risks through examples. But if we examine real examples of moderation, in many cases, the suggested conduct appears to contradict the rules. For example, in this group of slides about sexuality – which make no reference to the moderator's prerogative to conduct a contextual and personal analysis – some sexually-oriented nudity in digital art (a stick man with genitals, etc.) is allowed as long as there is not too much detail, while some hand-made pieces of classical art, such as the sculpture by Giambologna, the Rape of the Sabin Woman (1582) is forbidden. Since the sculpture does not contain any vulgarity, explicit sex, or sexually oriented nudity, we might think that the insistence on removal is due to the detailed bosom. We deem the application of these rules to again be very puzzling and ill-conceived.

4.4 Linguistic/semiotic issues

In the slides on terrorism,[41] we see five photos, each with three different captions.

The photos represent ISIS troops, some composed of children, ready to fight, taking very menacing stances; a photo showing the torture of four "infidels" hanging from crucifixes (plus one lying dead on the ground). The slides say that the moderator should delete these photos only when the captions express «support, praise or representation» of the image, while ignoring posts with «neutral» or «condemning» commentaries. The only exceptions to this rule are some Symbols or Leaders of primary focus (e.g., Bin Laden, a swastika, ETA's symbol, portraits of major leaders of the Jihad); the images should also, however, also be deleted if they appear without captions or in decontextualized uses. These rules of thumb have some intuitive validity but incur two major problems.

First of all, the definition of a neutral commentary is not clear. While we might be able to think of a journalistic description for the last three pictures,[42] the first two are hardly neutral: the first represents an execution squad killing some prisoners on a gallows, the caption «more deaths» does not make it clear whether it refers to producing a weapons of mass extermination or is a simple journalistic account (in any case, the picture is rather explicit); the second is definitively puzzling, since the caption «ISIS show of force creates fear» refers to the creation of fears generated by a battalion of soldiers, suggesting an idea of fear not inherently present in the picture. The caption for the third photo does not contain praise but instead depicts the crucifixion of four men: even without a caption, it will certainly be disturbing for many people.

The second problem concerns the condemning commentaries, some contain hate speech vocabulary (bastards, rats, animals, morons, and other expressions that are even worse). This vocabulary violates the FB community's standards on hate language. FB's rules ask its employees to delete any references to terrorist organizations, but in this case, the rule overrule other rules and could be modified according to general temporary sentiments that prevail at a given moment. In fact, these slides end with two specific exceptions to rules managing the moderation policy in the days after the 2016 Nice (France) Truck Attack and the Istanbul Airport Attack; in those cases, even indirect references to the attack should be deleted and marked as terrorist activities. In this specific case, even pictures referring to or depicting death should be deleted as «cruel» or «disturbing». Here we can see how policies that do not make a clear distinction between public and private perspectives fail; the same confusion is present when defining how "disturbing" content may be.

The slides pertaining to hate speech and anti-migrant messages[43] are particularly interesting because of the linguistic analysis they offer. FB decided to pay special attention to attacks addressed to protect categories. In our opinion, a specific example is particularly relevant: the rules suggest the removal of a group entitled «I fucking hate Christians», but the same rules determine in the next slide that a group called «I hate Christianity» should be ignored; in the same slides, similar treatment is suggested for content on homosexuality. Aren't moderators supposed to protect groups and people at risk?

They are but, as specified in the following slide, they are protecting people, not ideas. Hence the same rule (on hate speech per se) would prescribe deletion if my hate is directed against protected people, but never do so if I attack the very reasons why these people need to be protected: their ideas. Moreover, by these rules, "protection" applies to social classes, but not to appearances (e.g. being blonde, brunette, short, tall, fat, thin, etc.) or political ideology.

In the name of free speech, migrants are considered a quasi-protected category: without violating rules that regulate other protected communities (e.g. Christians, Muslims, gay people etc.), anyone is free to talk about segregating or excluding migrants, that they should be fired from their jobs, curse them, or even call them «thieves, robbers or

filthy». Many people would consider the use of 'filthy' as a word with racial overtones, making use of this term a racist act and an utterance constituting hate speech; FB's slide nevertheless allows for this practice saying, «Filthy is an adjective not a noun, we consider this to be a description of their appearance, not of their nature». Even after an update of these rules on April 2016, expressly forbidding violence and dehumanizing generalizations against Protected Categories, generalizations referring to an alleged race-dependent predisposition to kill, rape or not pay taxes were considered admissible as long as they avoided hate speech.

A similarly ambiguous and permissive policy determines the use of explicit sexual language. FB poses limits on how many details of sexual behavior can be contained in users' posts. Even considering that in many pop-culture environments many apparently explicit and vulgar expressions have lost their strictly sexual semantic reference, nonetheless this slide suggesting which cases to ignore appears extremely tolerant and is not likely to protect the wellbeing of many users on the platform.[44] The only sexual language forbidden seems to be direct and detailed description of sexual acts, with the exception of those reported in a «humorous context, insulting, educational, or figurative speech».

This policy is confusing, because it has nothing to do with the protection of moral decency and does not aim to protect younger users of the platform; even if we consider it to be aimed at creating a comfortable ambience for users, it seems to represent a partisan, biased, and sometimes incoherent vision of offensive language and pictures.

Some of the cases presented in FB's slide on racism are examples of tacky irony, but nonetheless they are considered harmless and moderators are suggested to ignore them. For example, a picture containing a famous athletic gesture by the famous American wrestler, Hulk Hogan,[45] shows the wrestler tearing his shirt up as a demonstration of strength. The caption links this act to the racist reactions of a non-black father, presumably against interracial marriage, who becomes aware of his daughter's sexual preference for black guys («When you find out your daughter likes black guys»). FB did not recognize any explicit racist intent in this meme and considered it innocuous. The slides offer a better explanation for another decision to ignore a controversial picture:[46] the caption for a photo depicting a giant agricultural machine harvesting a cotton field picture reads: «Super Ultra Nigger 9000. A modern machine for a modern racist». The slide specifies that this is a tricky case, where the caption enables a mocking interpretation of the racist slur; in fact, in the absence of this caption, the image would be considered to violate rules.

These cases are not easy to manage because they involve irony; still, we can recognize when irony is used in bad taste and could be offensive to black users in the community. Moreover, even if we like this kind of irony and consider it harmless, like FB, we should compare these cases to the use of «filthy» as a description of a migrant's appearance discussed above – an offensive usage so naively underestimated by the platform. Such rules constitute, at least, a "double standard" or, more precisely, a many-faceted group of standards, lacking any stable or consistent principle, such that the rulebook as a whole may not deserve to make use of the word "standard".

4.5 Exploiting FB Guidebook criticalities: A thought experiment

We have already mentioned the potential dangers of CCM and how it has to manage several unresolved ethical problems. This section underscores the ways in which such rules can easily be circumvented and exploited. First of all, we should consider the structural weaknesses of A.I. driven moderation: automatic flagging procedures rely on machine learning and image recognition. Practically, this means that if a user wants to pub-

lish a nude picture which will be ignored by the algorithms, he simply has to know how to edit his picture to trick the pattern recognition code. For example, he can change the color of the nude skin in the photo slightly, in order to avoid nude skin recognition.

While this example might be considered a border-line case where a user tries to circumvent the rules of moderation, the entire process has so many weak points, that we can imagine a far more dangerous case. In our opinion, a thought experiment can exemplify the structural weaknesses of the CM process. Let's imagine that we want to create a group aimed at isolating and radicalizing some border-line individuals in order to convince them to commit a violent racist crime. The first thing to do would be to avoid a group name that easily signals a form of hate against people, one that expresses a racist concept, such as: «I hate black» or, more metaphorically, «Lighting out darkness». A brief study of FB's rules of moderation (plus their General Terms and Community Standards), tells us that we need to avoid using certain common words and expressions to elude human moderation. Therefore, exploiting the lack of context sensitivity and the suboptimal linguistic knowledge of the moderators, we choose to use indirect, figurative language, possibly explaining our usages to group members in documents linked as external resources, which they can download from an external site. If we carefully avoid admitting any members who would be likely to flag our contents, or who might use explicit and detailed language that reveals our intentions, our group could potentially avoid the human watchmen hired by the platform as well as the CP algorithm, which has been trained on previous cases.

Fortunately, this project is not so easy to accomplish. But it highlights two important features of the moderation process: the need for moderators with linguistic and semiotic competence; and the potential risks of letting users know all of the rules applied in moderation. The first recommendation would only require that CPs include tests by professional in the hiring process. CPs should also establish better articulated control processes and possibly reduce outsourced and poorly controlled, moderation practices. But, with regard to the second consideration, we have to acknowledge that if all the moderation rules are made public and include sufficient detail, any user could find a way to bypass these rules, exploiting exposed weaknesses in the system. In our opinion, CPs should increase the transparency of the structures governing moderation, the moderation process, and the skills required of moderators, and privately explain the practice of moderation only to specific content creators (journalists, experts, government agencies, etc.). In this way, both the rulebook and the complete dataset would remain secret, preserving necessary privacy; at the same time, the mechanisms would be more clearly described and there would be reliable assistance for users who want to avoid making choices that violate the rules. Moreover, the users would then consider themselves to be more involved in the process and feel that they belonged to a more fair and trustworthy community.

5 How does CCM influence user behavior?

Academics have paid a great deal of attention to the ways in which social networks and their algorithms influence users. There are three main effects discussed in the literature: the *epistemic bubble*, *filter bubble,* and *echo chambers*. Our thesis is that none of these phenomena completely explain the cognitive influence CCM has on users. An epistemic bubble,[47] the most general of these concepts, occurs when a network provides inadequate information coverage, through a process of exclusion by omission. This means that members of the community will not receive all the relevant evidence, nor be exposed to a balanced set of arguments on certain topics; the evident result is the presence of several well-known cognitive biases (*confirmation, attention*, etc.). Of course, this happens be-

cause our epistemic activity should not be radically isolated from diverse sources of information and our social interactions should be based on *trust.*[48] It also incurs a common risk when we are allowed to select our favorite subjects (the most common selective behavior in a CP environment): we indulge in *selective exposure.*[49] This general and deeply debated effect is slightly different from a *filter bubble.*[50] A filter bubble occurs when selective exposure is produced by the social network's own algorithms, which choose which users' content we will see first. This personal filtering is opaque and even the users aware of its existence cannot easily recalibrate.[51] Finally, *echo chambers*[52] produce a selective effect by reinforcing and amplifying the posts of active human users, who continuously produce contents able to discredit any thesis opposed to those sustained in their community of interest, especially those related to politically and socially relevant issues.

We saw in Section 4 that CCM can, with its cognitive ambiguities and its subservience to external motivations (company profits or legal and political pressures), produce cognitive distortions. A user, without particular ideological motives, who sees his content deleted without a clear explanation for which part of that content violates the rules and how, is likely to become disoriented. Over the long term, this kind of consistent monitoring could produce an effect on users' communicative behaviors and impair their freedom of speech. Moreover, Geo-blocking content is an act of censorship that has a strong effect on users' perception of reality. For a Russian citizen, Google Maps reinforces the notion that Russia enjoys fair and recognized political control over Crimea.

This kind of influence, which we can call a *moderation filter*, falls within the scope of the epistemic bubble and has a similar effect to the filter bubble – but with two fundamental differences. When users are in a filter bubble, typical of social networks, their activity still influences, in one way or another, the existence of the bubble. By contrast, in the moderation filter bubble, users are just passively subjected to the cognitive influences of CCM. Whether or not they trust the other members of the community, and no matter what they read or view online, CCM continues to silently manipulating the contents they see and the creators and users they interact with. This influence is imposed, in a top-down fashion, by CPs and national governments and its range extends to both content creators and platform users. For these reasons, users can and should ask for more transparency and for a fair and participative appeal process, even if this requires their time and active participation.

Aside from this important passive consequence of moderation filters, there is another unpredictable effect on user activities. When a content creator has been moderated, warned, or even banned from the platform, they will end up changing their behavior, possibly without even knowing the rules of moderation applied or how their content violates these rules. Meanwhile, expert users and members of organized communities, who constantly update their knowledge of rules and community standards, can circumvent the intervention of moderators, as shown in our thought experiment, and thereby effectively bypass CCM. These dynamics produce two active distortions: common users, when moderated (even if unjustly), tend to obey these obscure and sometimes distorted laws and to spread them through their groups of contacts in order to avoid being banned from the community; in contrast, organized and ideologically oriented users can reiterate and disseminate malicious content without losing any formal legitimacy, hence strengthening their *echo chamber*. In this way, many users who have to contend with mild instances of moderation are led to actively reinforce the potentially negative effects of CCM.

This outline of CCM reveals a process that is imposed, opaque, partial, practiced by unskilled and exploited people; such methodological chaos leads us to fear that the *moderation filter* imposes important cogni-

tive distortions on user behavior. This distortion combines with some of the previously identified cognitive effects in new and complex ways: on one hand, it enhances the CP-generated epistemic bubble, as well as the effects of the bubbles and echo chambers, especially in countries where strong ideological censorship interacts with CCM rules; on the other hand, it limits, in incoherent and opaque ways, the freedom of expression of many users, while not concretely and effectively combating the activities of racist and extremist groups or communities that exist to protect socially deviant behaviors.

6 Who watches the watchmen?

The legal aspects of CCM are numerous and complex. They relate to the legitimacy and considerable risks associated with censorship and raise key questions related to freedom of speech, but also contribute to further refining the concept of social responsibility. A specific analysis of all such aspects is beyond our scope and capacity. But we can emphasize some important considerations related to authorities that have a powerful influence on moderation procedures. As has been the case with regard to many issues related to informatics and the internet, the US has pioneered many of the now international laws; this is also the case in relation to liabilities for internet content. In 1996, Title V of the Communications Decency Act made providers and users of an interactive computer service who publish information provided by others, immune from liability. This has always been a key legal advantage for CPs. However, the times are about to change.

At this point, it would be natural to ask if national governments would prove to be better authorities than CPs with respect to CM. Again, a critical lack of data and transparency makes this question hard to answer. Nations that directly control their online censorship are not willing to share their data. However, interesting laws have just appeared on stage among "western countries".

After the various political scandals involving FB, several countries considered relatively free and tolerant of internet freedom have decided to revise their policies on privacy, hate speech, defamation, and terrorism. In Europe, a leading model is the German law, NetzDG.[53] This law, in place since 1 January 2018, holds social media platforms responsible for combating online speech deemed illegal under domestic law. Other countries, such as France, UK, and Italy are discussing and predict the adoption of similar laws. As a summary explanation of the intent of such laws, we can say that NetzDG holds CPs – who include more than 2 million users in German territories – responsible for any violation of 22 «statutes» related to restricting hate speech, defamation, terrorism, pornography, and tampering with evidence. Platforms now must remove these contents within 24 hours; if they fail to comply, they risk fines of up to €50 million. While it is contemplated that several forms of screening and process controls included in this law, yet, to date there have been no requests to demand platforms institute greater transparency.

Many associations and even CP stakeholders fear that a consequence of these national laws will be limitations on users' freedom of speech, right to appeal, and more generally, an increase in removals and banned content. A platform regulated by this law has to integrate it with their in-house moderation process, applying the 22 statues immediately after internal moderation. The available data reveals that in the first six months of NetzDG application,[54] Google and Twitter have recorded greater numbers of reported items and an impressive percentage of 93% removal of reported content. FB and Change.org (the other two CPs involved) had less than two thousand reported items, with FB involved in a smaller removal percentage than other CPs (76,4%). This is because FB decided to integrate the examples from NetzDG in a hard-to-reach webpage linked to its flagging tool, while the others just include the option to view these exam-

ples/instructions alongside their main moderation commands. In the second year, the overall numbers of moderation episodes across all CPs has decreased, but the differential data for different CPs remains the same.

NetzDG had a significant effect on total reported and deleted items, especially with regard to hate speech and defamation. A number of political events and controversies in the German political context have influenced and continue to influence German public opinion and debates on this law. From this situation, we can inductively say that NetzDG increased the overall numbers of removals, especially those related to the inappropriate content targeted by the law (hate speech and defamation); but it is also possible that this happened because moderators and CPs were intimidated by the significant fines they can incur. Unfortunately, this case does not tell us if more democratic participation in the CM process leads to greater awareness and fairer CM practice. In effect, Germany wrote the rules but forced CPs to be their police and judge, changing the concept of moderation authority. What has happened more recently in the EU is even more interesting.

The Austrian politician, Eva Glawischnig-Piesczek, sued FB for not removing materials that contained false statements and offensive comments. This long international trial ended on October the 3rd 2019, when a verdict issued by the Court of Justice of the European Union included a landmark decision to consistently revisit the understanding of general monitoring obligations of hosting providers (Directive 2000/31 on electronic commerce). The verdict included the stipulation that FB could be forced to globally remove a post by any national court within the European Union's 28-member bloc if such content was determined to be defamatory or otherwise illegal. This decision cannot be appealed and gives European countries the power to apply takedown requests internationally and not just within their territory (Geo-blocking). Naturally, many scholars and organizations have analyzed this stipulation, expressing many concerns for freedom of speech, and fears of a generalized increase in arbitrarily moderated contents; they have also questioned what other laws governments could force CPs to comply with. One of the main dangers is that this verdict would lead CCM to make even stronger use of automation in order to address the global scope of removal. Moreover, the Court of Justice says that the CP can «be ordered to seek and identify the information equivalent to that characterized as illegal» and that national courts are allowed to order preventive deletions of «information with an equivalent meaning». Some key definitions in the sentence are not sharply drawn and unambiguous and do not clearly define tasks and tools for content analysis. Finally, several freedom-of-speech organizations[55] have expressed major concerns over its effect on user freedoms. The outlined legal scenario turns out to be very complicated and far from achieving better results in term of freedom of speech and cognitive distortion.

7 Conclusions

CCM is a necessary activity that has to contend with many complex issues. In order to make the process more effective and less dangerous for users' freedom and their online behavior, protocols should be discussed and tested with selected users and competent organizations. In our opinion, CPs should allow specialist groups to access their moderation data in order to monitor freedom of speech and academics should be allowed to investigate the cognitive effects of moderation. With the amount of user-generated content on the web, fake news and users' profiling, CCM pose serious threats to individual and community freedoms. In its current form, moderation risks posing a less well-understood version of the same threats. Lastly, legal approaches to CCM should prioritize users' freedom of speech and consider the possible effects of known CM protocols on users' online activities, convictions and on collective cognitive assets.[56]

Notes

[1] Cfr. T. GILLESPIE, *Custodians of the internet*, Yale University Press, New Haven (CT) 2018.
[2] Cfr. S.T. ROBERTS, *Behind the screen. Content Moderation in the shadow of the social media*, Yale University Press, New Haven (CT) 2019.
[3] Cfr. J. PREECE, D. MALONEY-KRICHMAR, C. ABRAS, *History of online communities*, in: K. CHRISTENSEN, D. LEVINSON (eds.), *Encyclopedia of community: From village to virtual world*, Sage Publications, Thousand Oaks 2003, pp. 1023-1027.
[4] Cfr. B. MCCULLOUGH, *How the internet happened*, Liveright Publishing, New York 2018.
[5] https://tools.ietf.org/html/rfc1855
[6] Cfr. G. KIRANOĞLU, *Copyright and the internet: The case of napster*, in: «International Journal of Human Sciences», vol. XIII, n. 2, 2016, pp. 2758-2767 - doi:10.14687/jhs.v13i2.3839.
[7] https://sciencenode.org/feature/How%20did%20smartphones%20evolve.php
[8] https://www.searchenginewatch.com/2014/07/08/mobile-now-exceeds-pc-the-biggest-shift-since-the-internet-began/
[9] Source: www.dataportal.com Data released by the major Social networks, update to January 2019.
[10] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/
[11] https://data.europa.eu/euodp/it/data/dataset/S2183_464_ENG
[12] Cfr. J. HINDS, E.J. WILLIAMS, A.N. JOINSON, *"It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge Analytica scandal*, in: «International Journal of Human-Computer Studies», vol. CXLIII, 2020 - doi:10.1016/j.ijhcs.2020.102498.
[13] https://www.nytimes.com/2018/12/27/world/facebook-moderators.html
[14] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[15] https://www.vice.com/en_us/article/7x478b/facebook-revenge-porn-nudes
[16] https://www.bbc.com/news/technology-51954968
[17] https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works
[18] *ibidem*
[19] https://www.theguardian.com/news/series/facebook-files
[20] www.facebook.com/communitystandards/
[21] Cfr. S.T. ROBERTS, *Behind the screen*, cit.
[22] http://gawker.com/5885714/inside-facebooks-outsourced-anti+porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads This article from 2014 is not available on Gawker's site, but the previous address can be reached through Wayback Machine (web.archive.org).
[23] https://www.facebook.com/terms.php
[24] https://uk.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUKKCN1GO2PN
[25] https://www.vox.com/2017/2/16/14632726/mark-zuckerberg-facebook-manifesto-fake-news-terrorism
[26] Cfr. S.T. ROBERTS, *Behind the screen*, cit., pp. 149-155 and p. 213.
[27] http://www.mcclatchydc.com/news/nation-world/national/article125953194.html
[28] https://www.theguardian.com/news/2017/may/21/facebook-moderators-quick-guide-job-challenges; https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content
[29] https://www.theguardian.com/news/series/facebook-files
[30] https://www.theguardian.com/news/gallery/2017/may/21/facebook-rules-on-showing-cruelty-to-animals
[31] https://www.theguardian.com/news/gallery/2017/may/24/how-facebook-handles-holocaust-denial
[32] https://web.archive.org/web/20110517092311/http://nation.com.pk/pakistan-news-newspaper-daily-english-online/Lahore/14-May-2011/LHC-orders-ban-on-Facebook-over-caricatures; https://en.wikipedia.org/wiki/Everybody_Draw_Mohammed_Day
[33] Cfr. T. GILLESPIE, *Custodians of the internet*, cit.
[34] https://mashable.com/2014/04/11/crimea-google/?europe=true
[35] https://www.theguardian.com/news/gallery/2017/may/22/sex-and-nudity-in-art-see-facebooks-rules
[36] http://www.reuters.com/article/usfacebook-content-insight-idUSKCN12S0D3
[37] https://www.theguardian.com/news/gallery/2017/may/22/how-sexual-activity-is-policed-on-facebook
[38] Picture available at: https://www.theguardian.com/news/gallery/2017/may/22/how-sexual-activity-is-policed-on-facebook#img-3
[39] Picture available at: https://www.theguardian.

com/lifeandstyle/shortcuts/2015/jul/07/instagram-facebook-female-nipple-ban-use-male-nipples-instead

[40] https://observer.com/2017/05/facebook-documents-reveal-policy-on-sex-nudity-art/

[41] https://www.theguardian.com/news/gallery/2017/may/24/how-facebook-guides-moderators-on-terrorist-content

[42] Picture available at: https://www.theguardian.com/news/gallery/2017/may/24/how-facebook-guides-moderators-on-terrorist-content#img-2

[43] https://www.theguardian.com/news/gallery/2017/may/24/hate-speech-and-anti-migrant-posts-facebooks-rules

[44] The slide defining "sexually explicit language" to ignore is available at: https://www.theguardian.com/news/gallery/2017/may/22/how-sexual-activity-is-policed-on-facebook

[45] Photo available at: https://www.theguardian.com/news/2017/may/21/ignore-or-delete-could-you-be-a-facebook-moderator-quiz

[46] Slide available at: https://www.theguardian.com/news/gallery/2017/may/24/hate-speech-and-anti-migrant-posts-facebooks-rules#img-18

[47] Our view of the *epistemic bubble* is influenced by and refers to Thi Ngiyen's definition (cfr. C. THI NGUYEN, *Echo chambers and epistemic bubbles*, in: «Episteme», vol. XVII, n. 2, 2018, pp. 141-161-doi:10.1017/epi.2018.32). But we deem it important to note here that, in philosophical studies, there is another epistemically relevant (and earlier) definition of the epistemic bubble, which appears in the seminal work *Epistemic bubble*, by John Woods (cfr. J. WOODS, *Epistemic bubbles*, in: S. ARTEMOV, H. BAR-RINGER, A. GARCEZ, L. LAMB, J. WOODS (eds.), *We will show them: Essays in honour of Dov Gabbay*, vol. II, College Publications, London 2005, pp. 731-774). The general effect of cognitive embubblement happens when: «A cognitive agent X occupies an epistemic bubble precisely when he is unable to command the distinction between his thinking that he knows P and his knowing P». This concept stimulated many epistemological studies and several important studies on ignorance in online communities (cfr. B. MILLER, I. RECORD, *Justified belief in a digital age: On the epistemic implications of secret internet technologies*, in: «Episteme», vol. X, n. 2, 2013, pp.117-134). Other authors have also been very important for this work (cfr. S. ARFINI, T. BERTOLOTTI, L. MAGNANI, *The diffusion of ignorance in on-line communities*, in: «International Journal of Technoethics», vol. IX, n. 1, 2018, pp. 37-50). Since the cognitive effect of CCM turns out to be so specific and related to its platform, we chose to rely directly on Thi Nguyen's definitions, specifically intended for online communities.

[48] Cfr. J. HARDWIG, *The role of trust in knowledge*, in: «The Journal of Philosophy», vol. LXXXVIII, n. 12, 1991, pp. 693-708.

[49] Cfr. J.L. NELSON, J.G. WEBSTER, *The myth of partisan selective exposure: A portrait of the online political news audience*, in: «Social Media + Society», July-September, 2017, pp. 1-13.

[50] Cfr. E. PARISER, *The filter bubble: What the internet is hiding from you,* Penguin, London/New York 2011.

[51] Cfr. J. HINDS, E.J. WILLIAMS, A.N. JOINSON, *"It wouldn't happen to me"*, cit.

[52] Cfr. C. THI NGUYEN, *Echo chambers and epistemic bubbles*, cit. See also K.H. JAMIESON, J.N. CAPPELLA, *Echo chamber: Rush Limbaugh and the Conservative Media Establishment*, Oxford University Press, Oxford 2008.

[53] https://germanlawarchive.iuscomp.org/?p=1245

[54] https://ssrn.com/abstract=3300636

[55] https://cmpf.eui.eu/eva-glawischnig-piesczek-v-facebook-ireland-limited-a-new-layer-of-neutrality/

[56] All internet sites mentioned in this paper were last visited on August 10th, 2020.

References

ARFINI, S., BERTOLOTTI, T., MAGNANI, L. (2018). *The diffusion of ignorance in on-line communities*. In: «International Journal of Technoethics», vol. IX, n. 1, pp. 37-50.

GILLESPIE, T. (2018). *Custodians of the internet*, Yale University Press, New Haven (CT).

HARDWIG, J. (1991). *The role of trust in knowledge*. In: «The Journal of Philosophy», vol. LXXXVIII, n. 12, 1991, pp. 693-708.

HINDS, J., WILLIAMS, E.J., JOINSON, A.N. (2020). "*It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge Analytica scandal*. In: «International Journal of Human-Computer Studies», vol. CXLIII, 2020 – doi:10.1016/j.ijhcs.2020.102498.

JAMIESON, K.H., CAPPELLA, J.N. (2008). *Echo chamber: Rush Limbaugh and the conservative media establishment*, Oxford University Press, Oxford.

KIRANOĞLU, G. (2016). *Copyright and the internet: The case of Napster*. In: «International Journal of Human Sciences», vol. XIII, n. 2, pp. 2758-2767.

MCCULLOUGH, B. (2018). *How the internet happened*, Liveright Publishing Corp., New York.

MILLER, B., RECORD, I. (2013). *Justified belief in a digital age: On the epistemic implications of secret internet technologies*. In: «Episteme», vol. X, n. 2, pp. 117-134.

NELSON, J.L., WEBSTER, J.G. (2017). *The myth of partisan selective exposure: A portrait of the online political news audience*. In: «Social Media + Society», July/September, pp. 1-13.

PARISER, E. (2011). *The filter bubble: What the internet is hiding from you,* Penguin, London.

PREECE, J., MALONEY-KRICHMAR, D., ABRAS, C. (2003). *History of online communities*. In: K. CHRISTENSEN, D. LEVINSON (eds.), *Encyclopedia of community: From village to virtual world*, Sage Publications, Thousand Oaks, pp. 1023-1027.

ROBERTS, S.T. (2019). *Behind the screen. Content moderation in the shadow of the social media*, Yale University Press, New Haven (CT).

THI NGUYEN, C. (2018). *Echo chambers and epistemic bubbles*. In: «Episteme», vol. XVII, n. 2, 2018, pp. 141-161.

WOODS, J. (2005). *Epistemic bubbles*. In: S. ARTEMOV, H. BAR-RINGER, A. GARCEZ, L. LAMB, J. WOODS (eds.), *We will show them: Essays in honour of Dov Gabbay*, vol. II, College Publications, London, pp. 731-774.