Lead author: Ibrahim Sabra

Assistant Lecturer, Faculty of Law, British University in Egypt, Egypt

Co-author: Mostafa Elkadi

Junior Collaborator, International Institute of International Humanitarian Law, Italy

Public Sphere Distortion in the Age of Internet Giants: Regulatory Pathways toward the Implications of Automated Online Content Filtering

Correspondence:

Ibrahim Sabra, Faculty of Law, British University in Egypt, Egypt

DOI: https://www.doi.org/10.54873/jolets.v2i1.70

E-mail: Ibrahim.Sabra@bue.edu.eg and Elkadi_Mostafa@alumni.ceu.edu

Citation:

Ibrahim Sabra and Mostafa Elkadi, Public Sphere Distortion in the Age of Internet Giants: Regulatory Pathways toward the Implications of Automated Online Content Filtering, Journal of Law and Emerging Technologies, Volume 2, Issue 1, April 2022, p. 51-86

Public Sphere Distortion in the Age of Internet Giants: Regulatory Pathways toward the Implications of Automated Online Content Filtering

Ibrahim Sabra and Mostafa Elkadi

Abstract

In an era of information profusion, Internet giants play a key role in determining the content that individuals consume online. Social media platforms, for instance, claim that their automated filters can provide users with a personalised online experience and end internet chaos. However, these platforms today use automated filtering extensively to curate content disseminated online in an opaque way to their users. Some believe that the negative impact of automated filtering is overstated since it empowers individuals to enjoy a tailored online experience based on their preferences; however, others argue that it has severe repercussions. This paper first sheds light on the Internet's role in reshaping the future of the media sector and its role as a watchdog. Secondly, it discusses the so-called "Automated Online Content Filtering" and a number of correlated concepts. It then analyses using a socio-legal approach the related controversy and the consequential implications of employing automated filtering on public sphere. Finally, it comparatively explains the adopted regulatory measures and recommended steps to minimise the prejudice caused by these filters. The paper concludes that due to profit-based engagement optimisation which drives social media platforms to de-prioritise content likely to be less engaging, automated filtering may amplify biases and extremism, induce the proliferation of false news and inflammatory content, and exacerbate the manipulation of the electorate, algorithmic bias, and censorship. Thus, the international community must take concrete regulatory measures to mitigate such ramifications and sway Internet giants to adopt standards that would lead to a healthier digital public sphere.

Keywords: Internet Giants - Content Filtering - Automated Decision-Making - Fake News - Algorithmic Bias - Filter Bubbles - Echo Chambers - Digital Rights - Freedom of Information - Public Sphere.

تشوه المجال العام في عصر عمالقة الإنترنت: المسارات التنظيمية تجاه الآثار المترتبة على التصفية الآلية للمحتوى عبر الإنترنت

إبراهيم صبره ومصطفى القاضي

الملخص

يلعب عمالقة الإنترنت في ظل فيض المعلومات الحالي دورًا رئيسيًا في تحديد المحتوى الذي يستهلكه الأفراد عبر الإنترنت. فعلى سبيل المثال، تدعي منصات وسائل التواصل الاجتماعي أن عوامل "التصفية الآلية" الخاصة بها يمكن أن توفر للمستخدمين تجربة شخصية في العالم الرقمي والمساعدة على إنهاء فوضى المعلومات على الإنترنت. ومع ذلك، تستخدم اليوم تلك المنصات "التصفية الآلية" على نطاق واسع لتنسيق وتنظيم المحتوى المنشور عبر الإنترنت بطريقة مبهمة لمستخدميها. يعتقد البعض أن الحديث عن التأثير السلبي للتصفية الآلية مبالغ فيه لأن الأخيرة تمكّن الأفراد من الاستمتاع بتجربة مصممة لهم عبر الإنترنت بناءً على تفضيلاتهم، ولكن يجادل آخرون بأن لها تداعيات خطيرة.

تُسلط هذه الورقة الضوء أولاً على دور الإنترنت في إعادة تشكيل مستقبل قطاع الإعلام ودوره الرقابي. ثانيًا، تناقش ما يسمى بوسائل "التصفية الآلية للمحتوى عبر الإنترنت" وعددًا من المفاهيم المرتبطة. ومن ثم تستخدم المنهج الاجتماعي القانوني لتحليل الجدل ذي الصلة والآثار المترتبة لاستخدام تلك الوسائل على الخطاب العام. أخيرًا، تشرح بشكل مقارن الإجراءات التنظيمية المعتمدة حاليًا والخطوات الموصى بها لتقليل الضرر الناجم عن "التصفية الآلية" للمحتوى الرقمي. تخلص الورقة إلى أنه نتيجة لمبدأ "تعزيز التفاعل الربحي" والذي يدفع منصات وسائل التواصل الاجتماعي إلى عدم إعطاء الأولوية للمحتوى الذي يُرجح أن يكون أقل جذبًا لانتباه المستخدمين، فإن "التصفية الآلية" قد تضخم التحيزات السياسية والتطرف، وتساعد على انتشار الأخبار الكاذبة والمحتوى التحريضي وتؤدي إلى تفاقم التلاعب بالناخبين والتحيز الخوارزمي والرقابة على المحتوى. لذلك، يجب على المجتمع الدولي اتخاذ تدابير تنظيمية ملموسة للتخفيف من آثار تلك التداعيات، وكذلك الضغط على عمالقة الإنترنت لاعتماد معايير من شأنها أن تؤدي إلى مجال عام رقمي أكثر صحة.

الكلمات الرئيسية: عمالقة الإنترنت – تصفية المحتوى – الاتخاذ الآلي للقرار – الأخبار الكاذبة – التحيز الخوارزمي – فقاعات التصفية – غرف الصدى – الحقوق الرقمية – حرية المعلومات – المجال العام.

Contents

"The opinions expressed in this publication are those of the authors and do not purport to reflect the opinions or views of their employers"

1- Bits of The Internet Giants Age

i. The Internet: A Game Changer

Media has always been recognised as the main inducement to public debate, acting as the "mediator" between citizens and politicians and serves the public interest by conducting reports, analyses and journalistic investigations.(1) It is proven that media consumption impacts the formation of individuals' identities, thus having a plural and free media promotes the "free-flowing of information and ideas in society", opens the door for an effective political dialogue and reflects diversified points of view.(2) That is why media is known as the "fourth estate" as it acts as a public watchdog and contributes to the well-functioning of democratic societies.(3) Accordingly, states throughout history had an obligation to maintain media freedom and plurality.(4)

The Internet brought tremendous change to the media sector leading to the emergence of digital media. The internet nearly eliminated all market barriers and facilitated easy access to various and wide-ranging sources of

(1) Rolph D and others, Media Law: Cases, Material and Commentary (2nd edn, Oxford University Press 2015); Vīķe-Freiberga V and others, "A Free and Pluralistic Media to Sustain European Democracy" (The Report of the High Level Group on Media Freedom and Pluralism 2013) <https://ec.europa.eu/information_society/media_taskforce/doc/pluralism/hlg/hlg_final_report.pdf>

(2) "How Media Can Be an Instrument of Peace in Conflict-Prone Settings" (UNDP Oslo Governance Centre 2017) <https://www.undp.org/content/dam/norway/undp-ogc/documents/UNDPOGC_Media_conflict round-table background paper.pdf>; Ligabo A, "PROMOTION AND PROTECTION OF ALL HUMAN RIGHTS, CIVIL, POLITICAL, ECONOMIC, SOCIAL AND CULTURAL RIGHTS, INCLUDING THE RIGHT TO DEVELOPMENT" (UNHRC 2008) <https://undocs.org/A/HRC/7/14>

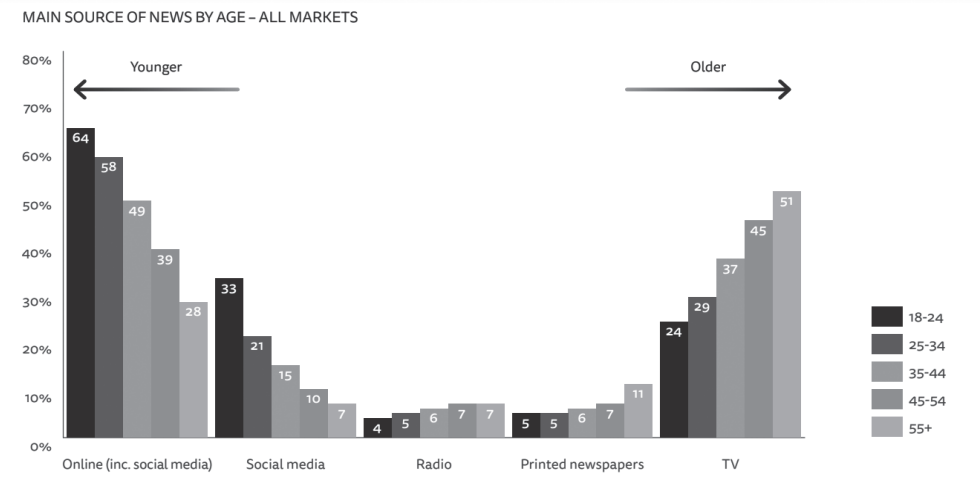
(3) "Declaration of Principles on Freedom of Expression" (IACHR 2000) <https://www.oas.org/en/iachr/mandate/Basics/declaration-principles-freedom-expression.pdf>; Price ME, Verhulst S and Morgan L, Routledge Handbook of Media Law (Routledge 2015) ch6; Schultz J, Reviving the Fourth Estate: Democracy, Accountability and the Media (Cambridge University Press 1998) 47–48; Finkelstein HR, "Report of the Independent Inquiry into the Media and Media Regulation" (Commonwealth of Australia 2012) <http://www.abc.net.au/mediawatch/transcripts/1205_finkelstein.pdf>; Liebes T, Curran J and Katz E, 'Public Sphere or Public Sphericules?' in Media, Ritual and Identity (Routledge 1998) 169; "Measuring Media Plurality" (Ofcom 2012) <https://www.ofcom.org.uk/__data/assets/pdf_file/0031/57694/measuring-media-plurality.pdf>

(4) "The Inter-American Legal Framework Regarding the Right to Freedom of Expression" (Office of the Special Rapporteur for Freedom of Expression Inter American Commission on Human Rights, OAS 2009) <http://www.oas.org/en/iachr/expression/docs/publications/INTER-AMERICAN LEGAL FRAMEWORK OF THE RIGHT TO FREEDOM OF EXPRESSION FINAL PORTADA.pdf>; Haraszti M, "Media Pluralism and Human Rights, Issue Discussion Paper" (Commission for Human Rights, Council of Europe 2011) <https://rm.coe.int/16806da515>; BÁRD P and BAYER J, "A Comparative Analysis of Media Freedom and Pluralism in the EU Member States" (Policy Department C: Citizens› Rights and Constitutional Affairs 2016) <http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571376/IPOL_STU(2016)571376_EN.pdf>

information.(1) It additionally facilitated, through social media platforms (SMPs), the rise of "User-Generated Content" and citizen journalism.(2)

Hence, with such numerous and appealing choices offered, people started seeing the internet as the main information provider and the most preferable public debate arena. Such a profound impact of the internet led media entities to actively get involved in the digital market. This was documented by a survey conducted by Reuters showing that news consumption, especially with younger age groups has notably shifted from printed media to online media (See figure 1).(3)

Figure 1



Nevertheless, the drastic increase of online content has caused an imbalance between the "limitless media" and the "limited attention" of internet users. Studies have shown that information overload overwhelms

(1) Vīķe-Freiberga V and others, "A Free and Pluralistic Media to Sustain European Democracy" (The Report of the High Level Group on Media Freedom and Pluralism 2013) <https://ec.europa.eu/information_society/media_taskforce/doc/pluralism/hlg/hlg_final_report.pdf>

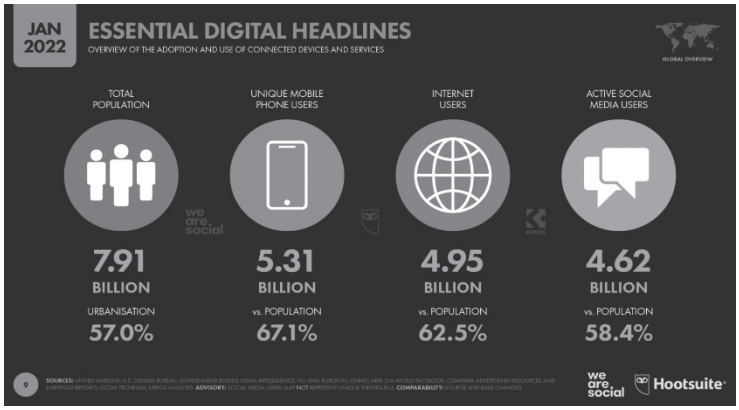
(2) Lessig L, Remix: Making Art and Commerce Thrive in the Hybrid Economy (Penguin 2008); McKay P, "Culture of The Future: Adapting Copyright Law to Accommodate Fan-Made Derivative Works in The Twenty-First Century" (2011) 24 Regent University Law Review 117; Pascu C and others, "Social Computing: Implications for the EU Innovation Landscape" (2008) 10 Foresight 37

(3) Newman N and others, "Digital News Report" (Reuters Institute for the Study of Journalism 2017) <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital News Report 2017 web_0.pdf>

individuals, prevents them from prioritising the relevant information, and eventually confuses them when processing content. Such massive amount of content available online has made it inevitable for people to refine what they get online and paved the way for content personalisation.(1)

Another key factor that has pushed towards content customisation is the growth of SMPs popularity and the flood of divergent and challenging content they brought especially when more than 90% of internet users are active on SMPs (See figures 2 and 3).(2) Consequently, the rise of filtering mechanisms has become inescapable.

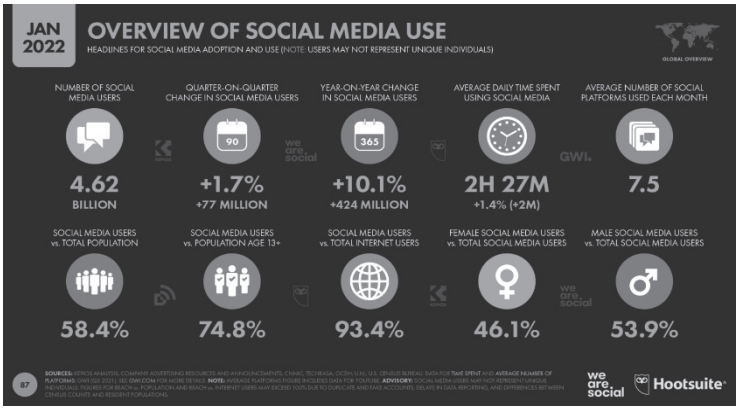
Figure 2



(1) "A Comparative Analysis of Media Freedom and Pluralism in the EU Member States" (Policy Department C: Citizens› Rights and Constitutional Affairs 2016) <http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571376/IPOL_STU(2016)571376_EN.pdf>; Eppler MJ and Mengis J, "The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines" (2004) 20 The Information Society 325; Leetaru K, "The Web Brought Us Limitless Information But Not The Tools To Manage It" (Forbes, May 26, 2019) <https://www.forbes.com/sites/kalevleetaru/2019/05/25/the-web-brought-us-limitless-information-but-not-the-tools-to-manage-it/>; Webster JG, Marketplace of Attention: How Audiences Take Shape in a Digital Age (MIT Press 2016)

(2) Kemp S, "DIGITAL 2022: GLOBAL OVERVIEW REPORT" (DATAREPORTAL, January 26, 2022) <https://datareportal.com/reports/digital-2022-global-overview-report>

Figure 3



ii. Automated Filtering: A Double-edged Sword

To properly understand the implications of automated filtering, it is important to shed light on a few concepts that are of significant impact on how filters function.

First, "Echo Chambers" which means that internet users voluntarily use online tools provided by SMPs to personalise what they get exposed to, allowing them to consume content that aligns with their beliefs and interact with users holding conforming views.(1)

Second, "Filter bubbles", which emanates from algorithmic curation that customises what users encounter online based on their personal data, browsing behaviour and online habits; hence, algorithms automatically exclude what seems irrelevant and recommend tailored content that best meets what users tend to interact and agree with.(2)

(1) Flaxman S, Goel S and Rao JM, "Filter Bubbles, Echo Chambers, and Online News Consumption" (2016) 80 Public Opinion Quarterly 298; Cardenal AS and others, "Digital Technologies and Selective Exposure: How Choice and Filter Bubbles Shape News Media Exposure" (2019) 24 The International Journal of Press/Politics 465; Dubois E and Blank G, "The Echo Chamber Is Overstated: the Moderating Effect of Political Interest and Diverse Media" (2018) 21 Information, Communication & Society 729

(2) Hannak A and others, "Measuring Personalization of Web Search" [2013] Proceedings of the 22nd international conference on World Wide Web - WWW 13; Cardenal AS and others, "Digital Technologies and Selective Exposure: How Choice and Filter Bubbles Shape News Media Exposure" (2019) 24 The International Journal of Press/Politics 465; Flaxman S, Goel S and Rao JM, "Filter Bubbles, Echo Chambers, and Online News Consumption" (2016) 80 Public Opinion Quarterly 298

Third, "Upload filters", this category of filters is mostly based on Artificial Intelligence (AI) empowered algorithms, and encompasses multiple acts, such as keyword filtering, or hash matching. Whereas the basic technology acts to remove or block certain types of speech either upon being uploaded or prior to being uploaded. This happens through having a digital fingerprint of the priorly removed content. The more advanced version of such filters are AI-driven algorithms that use machine learning through natural language processing and learn which kind of speech that is prohibited, then automatically removes such speech.(1)

Moving to the controversy over automated filtering, Cass R. Sunstein eloquently illustrates this dilemma when noted in his book that:

Technology has greatly increased people's ability to "filter" what they want to read, see, and hear. With the aid of the Internet, you are able to design your own newspapers and magazines…You need not come across topics and views that you have not sought out. Without any difficulty, you are able to see exactly what you want to see, no more and no less. If you are interested in politics, you may want to restrict yourself to certain points of view, by hearing only from people you like…Consumers…can design something very much like a communication universe of their own choosing.(2)

This statement reflects a viewpoint which believes that the internet personalisation empowers people's choices. Whether you agree or disagree with what it says, it is indisputable that the improper use of automated filtering could lead to unfavourable consequences.(3) The controversy thus stems from the current debate on the degree and extent of harm such filters could cause.

(1) Llansó E J No amount of "AI" in content moderation will solve filtering.s prior restraint problem' Big Data & Society January–June (2020) P. 2

(2) Sunstein CR, Republic.com 2.0 (Princeton University Press 2007) 3-5

(3) Vīķe-Freiberga V and others, "A Free and Pluralistic Media to Sustain European Democracy" (The Report of the High Level Group on Media Freedom and Pluralism 2013) 27; Sunstein CR, Republic.com 2.0 (Princeton University Press 2009) 7-12; "A Comparative Analysis of Media Freedom and Pluralism in the EU Member States" (Policy Department C: Citizens› Rights and Constitutional Affairs 2016) <http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571376/IPOL_STU(2016)571376_EN.pdf> 40-41

On the one hand, some scholars believe that the impact of automated filtering is overstated. They deem "technologies like web search and social networks reduce ideological segregation"(1) as such technologies increase choices, "exposure to diverse ideas"(2), "heterogeneous perspectives"(3) and help "break individuals free from insular consumption patterns"(4).

They also see that despite fragmentation and polarisation claims, "partisan news platforms are not the only or even the main sources of news and political information the general public rely on"(5) as statistics (see figure 4)(6) show that the majority still use general and impartial media outlets as a source of news(7) alongside the fact that "individuals are less likely to actively avoid information that contradicts their views"(8). Furthermore, they believe that most SMPs users get exposed to diversified political views, (9) and that because of SMPs' use of online filtering they are considered as the "least trusted medium" among different sources of news.(10)

(1) Flaxman S, Goel S and Rao JM, "Filter Bubbles, Echo Chambers, and Online News Consumption" (2016) 80 Public Opinion Quarterly 298

(2) Benkler Y, The Wealth of Networks: How Social Production Transforms Markets and Freedom (Yale University Press 2006)

(3) Messing S and Westwood SJ, "Selective Exposure in the Age of Social Media" (2012) 41 Communication Research 1042

Obendorf H and others, "Web Page Revisitation Revisited" [2007] Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 07 (٤)

(5) Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D., & Kleis Nielsen, R. (2017). Reuters Digital News Report 2017. Oxford: Reuters Institute for the Study of Journalism. Retrieved from http://www.digitalnewsreport.org/

(6) Dubois E and Blank G, "The Echo Chamber Is Overstated: the Moderating Effect of Political Interest and Diverse Media" (2018) 21 Information, Communication & Society 729

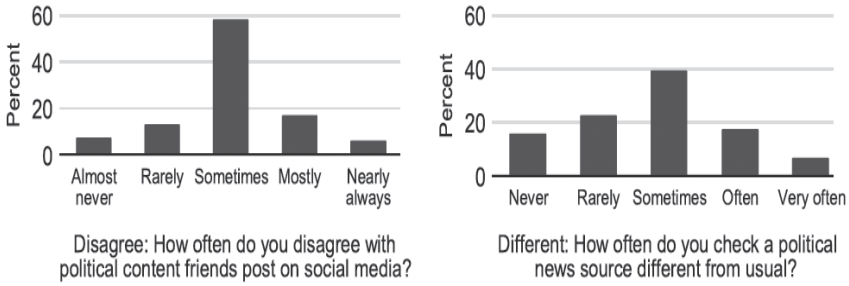
(7) Weeks BE, Ksiazek TB and Holbert RL, "Partisan Enclaves or Shared Media Experiences? A Network Approach to Understanding Citizens' Political News Environments" (2016) 60 Journal of Broadcasting & Electronic Media 248

(8) Garrett RK, "Echo Chambers Online?: Politically Motivated Selective Exposure among Internet News Users" (2009) 14 Journal of Computer-Mediated Communication 265

(9) Messing S and Westwood SJ, "Selective Exposure in the Age of Social Media" (2012) 41 Communication Research 1042

(10) Dutton W H and others, "Search and Politics: The Uses and Impacts of Search in Britain, France, Germany, Italy, Poland, Spain, and the United States" (2017) Quello Center Working Paper No. 5-1-17

Figure 4



On top of that, they heavily criticise "Single-Platform Studies" and find it "problematic" because seeking news from only one source is scarce to happen especially when living in a "high-choice media environment".(1) They backup this by explaining that Twitter, which was under intense scrutiny, is being used only by approximately 25% of the UK population that is "younger, wealthier, and better-educated than Britain as a whole" and therefore, can neither be regarded as an adequate representative segment of Britons nor their electoral votes.(2)

On the other hand, multiple experts, scholars, and activists perceive echo-chambers and filter bubbles as a serious issue significantly distorting the digital public sphere represented in SMPs, which they consider the main source of political information and news nowadays, and subsequently undermining the socio-political atmosphere and democracy. They believe that these mechanisms amplify political polarisation, online extremism, information pollution, and electoral manipulation. So, next, we will tackle each of these implications separately, however, it is important to mention that they are, to a large extent, interdependent.

(1) Dubois E and Blank G, "The Echo Chamber Is Overstated: the Moderating Effect of Political Interest and Diverse Media" (2018) 21 Information, Communication & Society 729

(2) Blank G and Lutz C, "Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google, and Instagram" (2017) 61 American Behavioral Scientist 741

There is an even bigger controversy when it comes to upload filters specifically. As it could be argued that the internet could be considered as "today's global marketplace of ideas",(1) and is the main source of political information as mentioned above, yet while this sentiment enhances the idea of freedom of expression, counter arguments arise claiming that the internet does not serve as the ideal marketplace of ideas, as theoretically, ideas that are "true claims, valuable ideas, and pro-social content" should prevail over "falsities, bad memes, and socially obnoxious content". However, this is not always the case, as the obnoxious content can result in impairing the rights of others. Therefore, it is the view of some that upload filters are a necessity.(2)

iii. The Ramifications of Automated Filtering

i. Political Polarisation

According to Sunstein,(3) a major danger to democracy is that individuals nowadays can easily customise their online experience by filtering what they get exposed to which is capable of leading to "the self-reinforcing spiral of polarisation".(4) He further clarifies that self-customisation is being escalated by algorithms which analyse users' behaviour to "unobtrusively adapt and provide the most relevant content without notifying the users"(5) and consequently, SMPs have become "the arbiters of what people see and what they don't"(6).

(1) Mouvement Raëlien Suisse v. Switzerland, App. No 16354/06 (ECtHR 13 July 2012) P. 54

(2) Sartor G and Loreggia A "The impact of algorithms for online content filtering or moderation: Upload Filters" (2020) A study requested by the European Parliament›s Committee on Citizens› Rights and Constitutional Affairs PE 657.101 P. 17-25

(3) A legal scholar and the Robert Walmsley University Professor at Harvard >https://hls.harvard.edu/faculty/directory/10871/Sunstein<

(4) Sunstein CR, Republic.com 2.0 (Princeton University Press 2007); Prior M, Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections (Cambridge University Press 2007); Bennett WL and Iyengar S, "A New Era of Minimal Effects? The Changing Foundations of Political Communication" (2008) 58 Journal of Communication 707

(5) Dylko I and others, "Impact of Customizability Technology on Political Polarization" (2017) 15 Journal of Information Technology & Politics 19

(6) Sunstein CR "Algorithms, Correcting Biases" (2018) Forthcoming, Social Research, SSRN < https://ssrn.com/abstract=3300171>

An example of the role of algorithms is what has been noted by Eli Pariser(1) that "Google keeps track of its user's choices and preferences, and changes its search results to suit them".(2) Several researchers(3) argue that both self-customisation and algorithms bias play a huge role in amplifying political polarisation within societies.(4) For instance, several studies have proven that filtering mechanisms "have increased exposure to pro-attitudinal political news articles and reduced exposure to counter-attitudinal political news articles"(5).

Moreover, in a study conducted on the most active political blogs during the US presidential election of 2004, it was found that these blogs primarily insert hyperlinks to "other blogs that align with their orientation" creating "their separate communities".(6)

Furthermore, in analyses of both Twitter and Facebook platforms, it has been demonstrated that users get exposed to content that is significantly in line with their views. In 2015 and 2016, statistics revealed that most of the politically motivated interactions on Twitter (around 75% of retweets) occur between users having similar political stances while on Facebook, information

(1) An online activist, chief executive of Upworthy and author of The Filter Bubble book >https://www.newamerica.org/our-people/eli-pariser/<

(2) Nguyen CT, "The Problem of Living inside Echo Chambers" (The Conversation, October 31, 2019) <https://theconversation.com/the-problem-of-living-inside-echo-chambers-110486>; Nguyen CT, "ECHO CHAMBERS AND EPISTEMIC BUBBLES" [2018] Episteme 1

(3) Stroud NJ, "Polarization and Partisan Selective Exposure" (2010) 60 Journal of Communication 556; Knobloch-Westerwick S, Choice and Preference in Media Use: Advances in Selective Exposure Theory and Research (1st edn, Routledge 2014)

(4) Dylko I and others, "Impact of Customizability Technology on Political Polarization" (2017) 15 Journal of Information Technology & Politics 19; Sîrbu A and others, "Algorithmic Bias Amplifies Opinion Fragmentation and Polarization: A Bounded Confidence Model" (2019) 14 Plos One

(5) Dylko I and others, "The Dark Side of Technology: An Experimental Investigation of the Influence of Customizability Technology on Online Political Selective Exposure" (2017) 73 Computers in Human Behavior 181; Murgia M, "Algorithms Drive Online Discrimination, Academic Warns" (Financial Times, December 12, 2019) <https://www.ft.com/content/bc959e8c-1b67-11ea-97df-cc63de1d73f4?fbclid=IwAR0S3PZGe0EyaHTY-uWYVldJr7ivJdniQvwflxfNLywKotUwFd_bWuOmbuM>

(6) Adamic LA and Glance N, "The Political Blogosphere and the 2004 U.S. Election" [2005] Proceedings of the 3rd international workshop on Link discovery - LinkKDD 05; Lawrence E, Sides J and Farrell H, "Self-Segregation or Deliberation? Blog Readership, Participation, and Polarization in American Politics" (2010) 8 Perspectives on Politics 141

about conspiracy theories "tends to spread within homogeneous and polarised communities".(1)

ii. Online Extremism

Unfortunately, extremism has spread all over the internet and SMPs were indeed the best place where extremist beliefs and related propaganda could be communicated.(2) Filtering mechanisms played a major role in the proliferation of extremist views on social media as many studies confirmed that individuals who take part in like-minded debates are more likely to adopt extreme positions compared to those who participate in diverse discussions.(3) That is why it is believed that such filters could make the internet a "breeding ground for extremism"(4) and "put social stability at risk"(5).

Robert Putnam,(6) shared his concerns that online filtration could help "white supremacists to narrow their circle to like-minded intimates".(7) Sunstein also thinks that such filtering process is the main reason behind "the rise in the number of hate groups and extremist organizations online".(8)

For example, research shows that the algorithm responsible for "Facebook's

(1) Barberá P and others, "Tweeting From Left to Right" (2015) 26 Psychological Science 1531; Vicario MD and others, "The Spreading of Misinformation Online" (2016) 113 Proceedings of the National Academy of Sciences 554; An J, Quercia D and Crowcroft J, "Partisan Sharing" [2014] Proceedings of the second edition of the ACM conference on Online social networks - COSN 14; Saez-Trumper D, Castillo C and Lalmas M, "Social Media News Communities" [2013] Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM 13

(2) Behr I and others, "Radicalisation in the Digital Era: The Use of the Internet in 15 Cases of Terrorism and Extremism" (RAND 2013) <https://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR453/RAND_RR453.pdf>

(3) Myers DG and Lamm H, "The Group Polarization Phenomenon." (1976) 83 Psychological Bulletin 602; Barberá P, "Social Media, Echo Chambers, and Political Polarization", (forthcoming. Social Media and Democracy: The State of the Field, edited by Nate Persily and Joshua Tucker. Cambridge University Press). 3 > http://pablobarbera.com/static/echo-chambers.pdf<

(4) Sunstein CR, Republic.com 2.0 (Princeton University Press 2007) 71

(5) Sunstein CR, Republic.com 2.0 (Princeton University Press 2007) 77

(6) A political scientist and a public policy professor at Harvard University

(7) Putnam RD, Bowling Alone: The Collapse and Revival of American Community (Simon & Schuster 2000) 178

(8) Barberá P, "Social Media, Echo Chambers, and Political Polarization", (forthcoming. Social Media and Democracy: The State of the Field, edited by Nate Persily and Joshua Tucker. Cambridge University Press). 3 > http://pablobarbera.com/static/echo-chambers.pdf<

Recommended Friends function" helped in connecting ISIL[1] members[2], while "Twitter's Who to Follow" algorithm suggested connecting to Islamist extremist's accounts in case the user followed al-Qaeda account.[3] Also, "YouTube's recommender system" prioritised far-right content when the user interacted with similar material.[4]

A further study on an emerging SMPs called Gab revealed that the latter became an echo-chamber for right-wing supporters. This led Gab to be known for its highly radical environment and instead of being "a social network that promotes free speech", it eventually ended up being a hate speech bubble and a haven for white supremacists.[5]

iii. Information Pollution

Although fake news is not a new concept as false claims have been regularly used to deceive and mislead the public since the time of ancient Greece, within the past two decades false information has gone viral worldwide in what is called the information pollution phenomenon.[6] The reason behind using "falsehoods" is the massive role it plays in "shaping people's attitudes toward a variety of high-profile political issues".[7] Unfortunately, with the rise of

(1) A terrorist organisation called Islamic State in Iraq and the Levant >https://www.britannica.com/topic/Islamic-State-in-Iraq-and-the-Levant<

(2) Waters G and Postings R, "Spiders of the Caliphate: Mapping the Islamic State's Global Support Network on Facebook" (Counter Extremism Project 2018) <https://www.counterextremism.com/sites/default/files/Spiders of the Caliphate (May 2018).pdf>

(3) Berger JM, "Zero Degrees of Al Qaeda" (Foreign Policy August 14, 2013) <https://foreignpolicy.com/2013/08/14/zero-degrees-of-al-qaeda/>

(4) Reed A and others, "Radical Filter Bubbles Social Media Personalisation Algorithms and Extremist Content" (Global Research Network on Terrorism and Technology 2019) <https://rusi.org/sites/default/files/20190726_grntt_paper_08_0.pdf>; Behr I and others, "Radicalisation in the Digital Era: The Use of the Internet in 15 Cases of Terrorism and Extremism" (RAND 2013) <https://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR453/RAND_RR453.pdf>; O'Callaghan D and others, "Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems" (2014) 33 Social Science Computer Review 459

(5) Lima L and others, "Inside the Right-Leaning Echo Chambers: Characterizing Gab, an Unmoderated Social System" [2018] 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

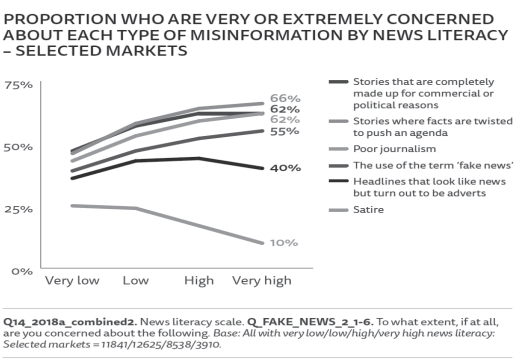
(6) Mara GM, The Civic Conversations of Thucydides and Plato: Classical Political Philosophy and the Limits of Democracy (SUNY Press 2008)

(7) Garrett RK, "The 'Echo Chamber' Distraction: Disinformation Campaigns Are the Problem, Not Audience Fragmentation" (2017) 6 Journal of Applied Research in Memory and Cognition 370

social media, widespread false information campaigns have been deployed for social, economic and political purposes. That is why the World Economic Forum has recognised false information as a major threat to the world.(1)

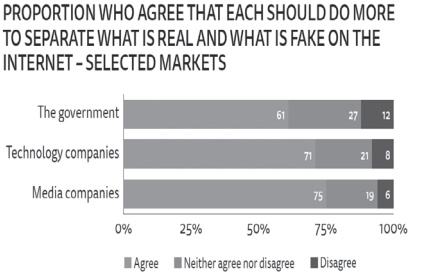
The Times mentioned that a fake story was once "shared at least 16,000 times on Twitter and more than 350,000 times on Facebook".(2) Buzzfeed also claimed that false news surpassed real news on Facebook during the months prior to the 2016 US elections.(3) These findings are reflected through surveys which show that people have become quite concerned about the spread of false information (See figures 5 and 6).(4) This tremendous issue raises a critical question about the reasons behind such pervasive spread of false information.

Figure 5



Q14_2018a_combined2. News literacy scale. **Q_FAKE_NEWS_2_1-6.** To what extent, if at all, are you concerned about the following. *Base: All with very low/low/high/very high news literacy: Selected markets = 11841/12625/8538/3910.*

Figure 6



Q_FAKE_NEWS_4_2_1-3. Please indicate your agreement with the following statements. Technology companies/media companies/the government should do more to make it easier to separate what is real and fake on the internet. *Base: Total sample: Selected markets = 46010.*

As previously explained, selective exposure to content creates insulated online islands besides negatively affecting content circulation.(5) Studies have shown that individuals' beliefs are hugely influenced by how their

(1) Howell L, "Digital Wildfires in a Hyperconnected World" (WEF Global Risks Report 2013) <http://www3.weforum.org/docs/WEF_GlobalRisks_Report_2013.pdf>

(2) Maheshwari S, "How Fake News Goes Viral: A Case Study" (The New York Times, November 20, 2016) <https://www.nytimes.com/2016/11/20/business/media/how-fake-news-spreads.html>

(3) Silverman C, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook" (BuzzFeed News, November 16, 2016) <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook>

(4) Newman N and others, "Digital News Report" (Reuters Institute for the Study of Journalism 2018) <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/digital-news-report-2018.pdf> 39 & 41

(5) Vicario MD and others, "The Spreading of Misinformation Online" (2016) 113 Proceedings of the National Academy of Sciences 554

communities approach events[1] especially when it comes to walled online communities "where users process information through a shared system of meaning", then adopt a uniform biased narrative usually in favour of their notion[2] and finally start sharing content aligning with such narrative while ignoring the rest.[3] This conclusion was consolidated by research showing that "an article shared by a trusted member of an individual's echo-chamber, but written by a source unknown to that individual, will be more likely to get consumed and shared than an article produced by a reputable news source but shared by someone viewed as less trustworthy".[4]

Furthermore, data explains that politically polarised individuals are "more likely to consume false news"[5] and "more resistant to counter speech that corrects that false news".[6] Deplorably, this undermines the quality of the information shared online causing a proliferation of "unsubstantiated rumours, mistrust and speculations"[7] and consequently, makes it difficult to either flag or correct false information[8]. In other words, echo-chambers make individuals more vulnerable to propaganda[9].

(1) Furedi F, Culture of Fear Revisited (2nd edn, Continuum International Publishing 2006)

(2) Bessi A and others, "Science vs Conspiracy: Collective Narratives in the Age of Misinformation" (2015) 10 Plos One; Mocanu D and others, "Collective Attention in the Age of (Mis)Information" (2015) 51 Computers in Human Behavior 1198; Bessi A and others, "Trend of Narratives in the Age of Misinformation" (2015) 10 Plos One; Zollo F and others, "Emotional Dynamics in the Age of Misinformation" (2015) 10 Plos One

(3) Vicario MD and others, "The Spreading of Misinformation Online" (2016) 113 Proceedings of the National Academy of Sciences 554

(4) Napoli PM, "What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble" (2018) 70 Federal Communications Law Journal 55; "‹Who Shared It?› How Americans Decide What News to Trust on Social Media" (American Press Institute, May 24, 2017) <https://www.americanpressinstitute.org/publications/reports/survey-research/trust-social-media/>

(5) Mocanu D and others, "Collective Attention in the Age of (Mis)Information" (2015) 51 Computers in Human Behavior 1198

(6) Napoli PM, "What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble" (2018) 70 Federal Communications Law Journal 55; Garrett RK, Weeks BE and Neo RL, "Driving a Wedge Between Evidence and Beliefs: How Online Ideological News Exposure Promotes Political Misperceptions" (2016) 21 Journal of Computer-Mediated Communication 331

(7) Difranzo D and Gloria-Garcia K, "Filter Bubbles and Fake News" (2017) 23 XRDS: Crossroads, The ACM Magazine for Students 32; Vicario MD and others, "The Spreading of Misinformation Online" (2016) 113 Proceedings of the National Academy of Sciences 554

(8) Bessi A and others, "Science vs Conspiracy: Collective Narratives in the Age of Misinformation" (2015) 10 Plos One

(9) Sunstein CR, #Republic: Divided Democracy in the Age of Social Media (Princeton University Press 2017)

Alongside echo chambers, the design of filtering technologies has influenced the media ecosystem, resulting in the dissemination and consumption of online news in a way that "undermined the production of legitimate news, while at the same time enhanced the production of false news".(1) For instance, a study conducted by Oxford University has concluded that fake news and propaganda shared on social networks are "supported by Facebook or Twitter's algorithms" by reducing the possibility of getting exposed to counter speech which mainly targets false information.(2) Thus, "the likelihood of fake news making it through the filter bubble increases. At the same time, the probability of legitimate news that counteracts that fake news decreases".(3)

iv. Electoral Manipulation

It was mentioned earlier that the implications of filtering mechanisms are interdependent. This could be best evidenced by addressing the topic of electoral manipulation especially after the surprising results of the 2016 US Presidential elections and the UK Brexit referendum as this topic reflects how such implications can collectively influence the voting outcome.

Neither Trump nor 'Leaving the EU' campaigns were expected to prevail as all indicators expected Hillary to win and UK to stay. That is why lots of people, even politicians and journalists, were shocked by the outcome. After some digging, experts started referring to filter bubbles and echo-chambers as the reason behind this, mainly because of the isolated communities they created alongside the fact that people who stay in these walled groups "fail to

(1) Napoli PM, "What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble" (2018) 70 Federal Communications Law Journal 55

(2) Hern A, "Facebook and Twitter Are Being Used to Manipulate Public Opinion – Report" (The Guardian, June 19, 2017) <https://www.theguardian.com/technology/2017/jun/19/social-media-proganda-manipulating-public-opinion-bots-accounts-facebook-twitter>

(3) Pariser E, The Filter Bubble: What the Internet Is Hiding from You (Penguin Books 2012); Bessi A and others, "Science vs Conspiracy: Collective Narratives in the Age of Misinformation" (2015) 10 Plos One; Napoli PM, "What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble" (2018) 70 Federal Communications Law Journal 55

evaluate alternative viewpoints, fail to consider risks, haze any members who disagree, and even ignore the possibility of a negative outcome".(1)

Shapiro(2), for example, has argued that search engine companies have the capacity to substantially influence elections results. He supported his claim by showing that "biased search rankings can shift the voting preferences of undecided voters by 20% or more". Shapiro also mentioned that while TV has the power to shift the results of elections, the internet has a much more significant impact. He mentioned that all information related to elections are ranked by algorithms which eventually favour some over the rest. Thus, algorithmic search rankings have become a new type of "social influence".(3) Additionally, political campaigns have used message targeting services such as "lookalike modelling", which is available through social media, to "largely target swing or undecided voters". Such a tactic "ultimately undermines the ability of voters to engage in meaningful democratic deliberation".(4)

What is alarming here is that SMPs are considered the primary source of news for around 61% of the youth, according to Pew Research.(5) It is more intriguing that 34% of internet users between the age of 18 and 34 said that the information and news they see on social media "influence their vote".(6) Hence, tampering news disseminated on SMPs can obviously lead to electoral manipulation.

Accordingly, countries had to interfere either directly by imposing

(1) How Filter Bubbles Distort Reality: Everything You Need to Know" (Farnam Street, November 14, 2019) <https://fs.blog/2017/07/filter-bubbles/>

(2) A professor of Politics and International Affairs at Princeton University and co-directs the Empirical Studies of Conflict Project >https://scholar.princeton.edu/jns/home<

(3) Epstein R and Robertson RE, "The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections" (2015) 112 Proceedings of the National Academy of Sciences

(4) Harker M, "Political Advertising Revisited: Digital Campaigning and Protecting Democratic Discourse" [2019] Legal Studies 1; "Internet and Electoral Campaigns: Study on the Use of Internet in Electoral Campaigns" (MSI-MED Council of Europe 2017) <https://rm.coe.int/use-of-internet-in-electoral-campaigns-/16807c0e24>

(5) Difranzo D and Gloria-Garcia K, "Filter Bubbles and Fake News" (2017) 23 XRDS: Crossroads, The ACM Magazine for Students 32; Mitchell AE, Gottfried JE and Matsa KE, "Facebook Top Source for Political News Among Millennials" (Pew Research Center›s Journalism Project, June 1, 2015) <https://www.journalism.org/2015/06/01/facebook-top-source-for-political-news-among-millennials/>

(6) "Internet and Electoral Campaigns: Study on the Use of Internet in Electoral Campaigns" (MSI-MED Council of Europe 2017) <https://rm.coe.int/use-of-internet-in-electoral-campaigns-/16807c0e24>

regulations or indirectly through pushing internet giants to implement self-regulatory measures, in order to minimise the negative impact of such serious issues on the flow of information and news online. The following section will address the current regulations and future recommendations.

v. Algorithmic Bias

While there are positive implications for the use of upload filters, such as their fast processing of data that otherwise would not be reviewed and the removal of copyright infringed materials,(1) yet there are also negative implications.

In the design of the upload filter, the biases of humans can be embedded within the algorithm. This impacts the behaviour of the system and can result in algorithmic bias while filtering content.(2) For example, the word "Mexican" was identified as a slur on the datasets of the SMP Instagram, because the word "Mexican" had an association with illegality.(3)

Those automated biases could also be reflected in speech that needs more context, such as cultural and social considerations, as there are different standards for speech in different communities, some speech may be acceptable in one place but not acceptable in the other. This corelates to the discussions that have been ongoing about the removal of content displaying female breasts in the context of breastfeeding or abstractly on SMPs. In this instance the filtering system could result in discriminatory judgements through the automation of content removal as "the content delivered by or concerning certain groups may be excluded or deprioritised", this also includes political manipulation in times of elections.(4)

(1) Kaye D, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 6; Sartor G and Loreggia A "The impact of algorithms for online content filtering or moderation: Upload Filters" (2020) A study requested by the European Parliament›s Committee on Citizens› Rights and Constitutional Affairs PE 657.101 P. 46-47

(2) Kaye D, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 6

(3) Kaye D, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 15

(4) Sartor G and Loreggia A "The impact of algorithms for online content filtering or moderation: Upload Filters" (2020) A study requested by the European Parliament›s Committee on Citizens› Rights and Constitutional Affairs PE 657.101 P. 46-47

The removal could happen instantly, or at a later time after sharing the content, all done through the algorithms of the upload filters. Thus, the mere concept of automating the removal of content has both positive as well as negative implications. The negative implications mainly encompass discriminatory removal of content.

vi. Censorship

Another issue arises from the automation of upload filters, more specifically the issue of censorship where the automated upload filters would remove content whenever and, in any manner, they deem fit. This specific issue was a concern that was voiced by the Special Rapporteur on freedom of expression arguing that

Users and civil society experts commonly express concern about the limited information available to those subject to content removal or account suspension or deactivation, or those reporting abuse such as misogynistic harassment and doxing. The lack of information creates an environment of secretive norms, inconsistent with the standards of clarity, specificity, and predictability. This interferes with the individual's ability to challenge content actions or follow up on content-related complaints; in practice, however, the lack of robust appeal mechanisms for content removals favors users who flag over those who post.(1)

As emphasised earlier, the automation has the positive implication of being able to assess huge amounts of content. Yet the problematic part is the removal process, as it may result in pre-publication censorship or over-blocking.(2) One example is when the kindergarten teacher Mr. Frédéric Durand-Baïssas uploaded a nude painting by Gustave Courbet. At the time, the painting did not go against Facebook's policy, yet it was removed and Mr.

(1) Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (6 April 2018) A/HRC/38/35, para 58.
(2) Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (6 April 2018) A/HRC/38/35, para 32-33.

Frédéric's account was suspended. It goes without saying that the original claim in the case was of censorship nature.(1)

What happened in the previous case is susceptible to happening at any time, as the machine learning algorithms, that most SMPs use for upload filters, can grow to be more unpredictable as they learn and adapt, as one of their capabilities is to identify new issues and solve them. It is logical to say that neither the new issue nor solving it would be foreseeable by humans.(2)

Looking at this implication from the users' perspective, the operation of those upload filters can be -and most likely is- known to the users. Yet the removal process is not obvious for the user with specific regards to the content that they share, the users are most likely not going to know before sharing their content whether it is going to be deemed offensive or appropriate, "this means that individuals will often have their expression rights adversely affected without being able to investigate or understand why, how or on what basis."(3) This is while also considering that SMPs "may change their own rules and algorithms over time."(4)

Thus, the implication of censorship is evident not only as a direct result of the automated filtering process but also indirectly by pushing users to impose self-censorship on their own out of fear that their content will be flagged or filtered by algorithms.

iii. A Robust Internet Regulatory Framework For A Healthier Public Sphere

i. Current Regulatory Framework

Almost all the measures adopted to restrict or regulate the usage of

(1) Bishara H, ‹Facebook Settles 8-Year Case With Teacher Who Posted Courbet'S "Origin Of The World"› (Hyperallergic, 2019) <https://hyperallergic.com/512428/facebook-settles-8-year-case-with-teacher-who-posted-courbets-origin-of-the-world/>

(2) Kaye D, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 8.

(3) Kaye D, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 31-32.

(4) Kaye D, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (29 August 2018) A/73/348, para 40

automated filtering were undertaken for the protection of freedom of expression and the free flow of information, which is enshrined in multiple international and regional conventions, such as Article 19 of the International Covenant on Civil and Political Rights and the Universal Declaration of Human Rights, Article 10 of the European Convention on Human Rights, Article 9 of the African Charter on Human and Peoples' Rights, and Article 13 of the American Convention on Human Rights (ACHR) which is the only international convention to directly prohibit almost all types of "censorship".(1)

While the phrasing of the Articles in such instruments are not inclusive enough to regulate how filters operate, the Articles agree that restrictions on either offline or online speech, arguably including automated online filtering, should conform to an accumulative three-step test that comprises of:

Legality: this requirement provides that the law must be accessible, foreseeable, and formulated with sufficient precision that would allow individuals to regulate their conduct accordingly and provide adequate safeguards against arbitrary misuse;(2)

Legitimacy: this requires that a restriction must pursue one of the legitimate aims exhaustively mentioned in Article 19(3) such as public order, national security, rights of others, etc.; and

Necessity: it stipulates that a restriction must be necessary in a democratic society which has two prongs; firstly, the restriction must be justified by a "pressing social need" that is related to the legitimate aim, and secondly the restriction must be "proportionate to the legitimate aim pursued".(3)

However, the issue in this area is that these conventions were directed to States, therefore they entail a State obligation,(4) not a corporate obligation.

(1) Article 13 (2) of the ACHR grants an exception in paragraph 4 by subjecting "public entertainment" to prior censorship when it is provided by law, for the sake of "moral protection of childhood and adolescence".
(2) Rekvényi v. Hungary App no 25390/94 (ECHR, 20 May 1999) para 34
(3) Jerusalem v. Austria App no 26958/95 (ECtHR, 27 Feb 2001), para.33; Dalban v. Romania App no 28114/95 (ECtHR, 28 Sept 1999), para.47
(4) UN Human Rights Committee (HRC), General comment no. 31 [80], The nature of the general legal obligation imposed on States Parties to the Covenant, 26 May 2004, CCPR/C/21/Rev.1/Add.13, Para. 2&6

Consequently, this results in having courts assessing the restrictions imposed by SMPs on online content through applying the aforementioned cumulative three-part test.(1)

In 2010, the Special Rapporteurs in a joint report listed challenges to freedom of expression and stated their concern that the free flow of information online could be hindered.(2) Nine years later, the same Special Rapporteurs identified the issue that private control over SMPs can infringe upon freedom of expression and access to information, as corporations now have "enormous power" in regulating the flow of information. (3)

The Special Rapporteurs urged States to take actions such as the adoption of policies for online platforms in order to regulate content, and thereby hold companies responsible for violations of human rights, specifically in compliance with the Guiding Principles on Business and Human Rights.(4) This Guiding Principles, although not binding, is one of the most well-established soft law documents in international human rights law and is widely upheld by international corporations, including internet giants. It constitutes a sound ground to prevent abuse by corporations as it sets clear, adequate standards that draw corporate responsibility to respect human rights and ensure the right to access remedies.(5)

For example, it sets out foundational principles for business enterprises to adhere to such as "avoid[ing] causing or contributing to adverse human rights impacts" and "seek[ing] to prevent or mitigate [such] impacts that are directly

(1) Delfi AS v. Estonia GC App No 64569/09 (ECtHR 16 June 2015)
(2) OAS, ‹TENTH ANNIVERSARY JOINT DECLARATION: TEN KEY CHALLENGES TO FREEDOM OF EXPRESSION IN THE NEXT DECADE› (Oas.org, 2010) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=784&lID=1>
(3) OAS, ‹TWENTIETH ANNIVERSARY OF THE JOINT DECLARATION: CHALLENGES TO FREEDOM OF EXPRESSION IN THE NEXT DECADE› (Oas.org, 2019) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=1146&lID=1>
(4) OAS, ‹TWENTIETH ANNIVERSARY OF THE JOINT DECLARATION: CHALLENGES TO FREEDOM OF EXPRESSION IN THE NEXT DECADE› (Oas.org, 2019) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=1146&lID=1>
(5) "Guiding Principles on Business and Human Rights" (United Nations 16 June 2011) <https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf>

linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts",(1) while also adopting "a policy commitment to meet their responsibility to respect human rights",(2) with a "complaint mechanisms" put into place that would afford "suitable remediation" in case of abuse.(3)

It was further stated in a Joint Declaration in 2011 that "Self-regulation can be an effective tool in redressing harmful speech".(4) Thus, promoting the idea of self-regulation. However, in the same Joint Declaration, it was mentioned that "Content filtering systems which are imposed by a government or commercial service provider and which are not end-user controlled are a form of prior censorship and are not justifiable as a restriction".(5) This can be contradictory to the previous statement that promoted self-regulation, but "filtration or blocking should be designed and applied so as to exclusively impact the illegal content without affecting other content".(6) In that sense there would be no violation to the freedom to receive or impart information.

It is noteworthy to mention that despite States' push for SMPs to implement a self-regulatory mechanism that would actively monitor content published on their platforms, the EU E-Commerce Directive removed any general monitoring obligation on online platforms to monitor user generated content, as it stated in Article 15 (1) that "Member States shall not impose a general obligation on providers… to monitor the information which they transmit or store, nor a general obligation actively to seek facts or circumstances indicating

(1) "Guiding Principles on Business and Human Rights" (United Nations 16 June 2011) Principle 13 <https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf>
(2) "Guiding Principles on Business and Human Rights" (United Nations 16 June 2011) Principle 15 <https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf>
(3) "Guiding Principles on Business and Human Rights" (United Nations 16 June 2011) Principles 22, 29, & 31 <https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf>
(4) OAS, ‹JOINT DECLARATION ON FREEDOM OF EXPRESSION AND THE INTERNET› (Oas.org, 2011) Para 1 (E) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1>
(5) OAS, ‹JOINT DECLARATION ON FREEDOM OF EXPRESSION AND THE INTERNET› (Oas.org, 2011) Para 1 (B) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1>
(6) Marino C B 'Freedom of Expression and the Internet' Office of the Special Rapporteur for Freedom of Expression IACHR (2013) CIDH/RELE/INF. 11/13, P. 37

illegal activity."[1] Thus, proactively monitoring content is not compulsory, and would arguably violate the freedom to receive or impart information.

This was convincingly affirmed by the Court of Justice of the EU ("CJEU") where the court ruled that online platforms are not obligated by any means to install filtering systems.[2] However, the same principle was not upheld in the European Court of Human Rights ("ECtHR"), whereas Delfi, an online platform was held liable for hosting hateful content for different reasons on top of which comes the fact that its filters did not operate in a manner that allowed for filtering unsophisticated hateful content that included "manifest expressions of hatred and blatant threats to the physical integrity" which is not protected speech under Article 10 of the European Convention on Human Rights (ECHR). The Court thus concluded that due to "this failure of the filtering mechanism these clearly unlawful comments remained online for six weeks and consequently "limited [Delfi] ability to expeditiously remove the offending comments".[3]

Furthermore, numerous actions were taken by the EU to achieve a healthier internet. Regarding disinformation diffusion, several online platforms and the EU have signed a self-regulatory code of practice addressing the spread of online disinformation.[4] The code promotes more transparent political advertising, the banning of fake accounts and the "demonetisation of purveyors of disinformation".[5] Moreover, a group of experts was set to assess policy initiatives related to countering misinformation online and present an analysis alongside recommendations in a detailed report.[6] Also, based on the

(1) Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (‹Directive on electronic commerce›) [2000] OJ L 178, art 15

(2) Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM) Case C-70/10 (CJEU 24 November 2011) Para 53; Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV Case C-360/10 (CJEU 16 February 2012) para 52

(3) Delfi AS v Estonia [GC] App no 40287/98 (ECtHR 16 June 2015) para 156

(4) "Code of Practice on Disinformation" (Digital Single Market - European Commission, June 17, 2019) <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation>

(5) "Code of Practice on Disinformation" (Digital Single Market - European Commission, June 17, 2019) <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation>

(6) "Tackling Online Disinformation" (Digital Single Market - European Commission, September 13, 2019) <https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation>

European Commission communication on tackling online disinformation,(1) an independent European Network of Fact-Checkers has been launched called SOMA to conduct fact-checking activities and provide online tools to fight disinformation online.

When it comes to online extremism, the EU has signed a voluntary code of conduct with some internet giants to encounter the spread of hateful content online.(2) The signatories agreed to assess the majority of flagged content within 24 hours and act accordingly, set community standards which prohibit hate speech on their platforms, publish annual transparency reports, "establish national contact points to communicate with competent national authorities", and enlarge their cooperation with civil society.(3) Thanks to the code, SMPs are currently reviewing 89% of the reported content within 24 hours and 72% of the illegal hateful content is removed, compared to "40% and 28% respectively when the code was first launched".(4)

Moving to the UK, a white paper(5) was recently introduced setting out an action plan to encounter the spread of harmful content online.(6) According to it, a new "statutory duty of care" will be established to push tech companies to be more responsible and active when dealing with harmful content disseminated on their platforms. An independent regulator will oversee these companies and their compliance with this duty of care and in case of breaching the latter,

(1) "COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS, Tackling Online Disinformation: a European Approach" (European Commission 2018) <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0236&from=EN>

(2) "Code of Conduct on Countering Illegal Hate Speech Online" (European Commission 2016) <https://ec.europa.eu/info/sites/info/files/code_of_conduct_on_countering_illegal_hate_speech_online_en.pdf>

(3) Jourová V, "How the Code of Conduct Helped Countering Illegal Hate Speech Online" (DG JUST, European Commission 2019) <https://ec.europa.eu/info/sites/info/files/hatespeech_infographic3_web.pdf>

(4) "Countering Illegal Hate Speech Online – EU Code of Conduct Ensures Swift Response" (European Commission, February 4, 2019) <https://ec.europa.eu/commission/presscorner/detail/en/IP_19_805>

(5) White papers are policy documents produced by the Government that set out their proposals for future legislation. Please, visit the UK Parliament for further information: https://www.parliament.uk/site-information/glossary/white-paper/

(6) Home Office news team, "Home Office in the Media Blog: Monday 18 February" (GOV.UK blogs, February 18, 2019) <https://homeofficemedia.blog.gov.uk/2019/02/18/home-office-in-the-media-blog-monday-18-february/>

the regulator will be empowered to impose "substantial fines" and "liability on the senior management".

The framework includes "setting codes of practice relating to illegal harms", "the regulator's power to require annual transparency reports", encouraging companies to "improve the ability of researchers to access their data" and companies' adoption of user-friendly complaints system with a reasonable timeframe for responding and taking actions. The paper, further, urged companies to "invest in the development of safety technologies" to guarantee a safe experience for their users. It also explained that a new "Online Media Literacy Strategy" will be arranged by the government to raise awareness through education.(1)

As for other countries, several have promulgated "Anti Fake News and Cybercrime Laws" in order to combat the spread of harmful content online such as Germany,(2) China, Malaysia, Egypt and Kenya, however, some of these laws have been heavily criticised for allegedly being abusively restricting online freedom of expression.(3) In Kenya, for example, the "Computer Misuse and Cyber Crimes Act", has been challenged before the Kenyan High Court and got temporarily frozen.(4) But, in 2020, the High Court of Kenya dismissed the motion and upheld the Act as constitutional which received serious censure from various human rights organisations.(5)

(1) "Online Harms White Paper" (HR Government 2019) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf>

(2) "Germany Approves Plans to Fine Social Media Firms up to €50m" (The Guardian, June 30, 2017) <https://www.theguardian.com/media/2017/jun/30/germany-approves-plans-to-fine-social-media-firms-up-to-50m>

(3) "Initiatives to Counter Fake News in Selected Countries" (The Law Library of Congress, Global Legal Research Directorate 2019) <https://www.loc.gov/law/help/fake-news/counter-fake-news.pdf>; Harriss L and Raymer K, "Online Information and Fake News" (POST, UK Parliament, July 12, 2017) <https://researchbriefings.parliament.uk/ResearchBriefing/Summary/POST-PN-0559>

(4) "Petition 206 of 2018" (Kenya Law 2018) <http://kenyalaw.org/caselaw/cases/view/159286>; Kakah M, "Court Suspends Portions of Cybercrime Law" (Daily Nation, June 28, 2020) <https://www.nation.co.ke/news/Court-suspends-portions-of-cybercrime-law/1056-4585936-thh4s5/index.html>

(5) "Kenya: Court of Appeal's ruling strikes further blow to free expression and privacy" (Article 19, August 14, 2020) <https://www.article19.org/resources/kenya-court-of-appeals-ruling/#:~:text=On%207%20August%202020%2C%20a,2018%20on%2020%20February%202020.>

Some other countries (e.g., Argentina,[1] Canada[2] and Czech Republic[3]) have set up units that will be in charge of detecting disinformation attempts, labelling fake news, and preventing any foreign disinformation campaigns, particularly during elections.[4]

ii . Proposed Regulatory Measures

Many initiatives have been proposed to tackle the abovementioned issues including the allocation of funds for research on the change of media environment in order to "provide a solid academic basis for policy initiatives in this field" and allowing more involvement by the civil society in the online media regulatory process.[5]

In conjunction with the developments in the field of automated filtering empowered by AI, stakeholders and parties that may be affected by the functioning of such systems should familiarise themselves with the process of how the system makes its decisions, and its consequences so they would be able to make an informed decision as to whether they would engage with the system.[6]

From a legal perspective, appropriate laws should be enacted to regulate political advertising online to set clear accountability rules, deterring sanctions,

(1) Robinson L, "Fake News Persists in Argentina as Election Draws Near" (Buenos Aires Times, September 28, 2019) <https://www.batimes.com.ar/news/argentina/fake-news-persists-in-argentina-as-election-draws-near.phtml>

(2) Funke D and Flamini D, "A Guide to Anti-Misinformation Actions around the World" (Poynter) <https://www.poynter.org/ifcn/ anti-misinformation-actions/>

(3) "Centre Against Terrorism and Hybrid Threats" (Ministry of the Interior) <https://www.mvcr.cz/cthh/clanek/centre-against-terrorism-and-hybrid-threats.aspx>; Goodman E, "How Has Media Policy Responded to Fake News?" (Media@LSE, February 7, 2017) <https://blogs.lse.ac.uk/medialse/2017/02/07/how-has-media-policy-responded-to-fake-news/>

(4) "Initiatives to Counter Fake News in Selected Countries" (The Law Library of Congress, Global Legal Research Directorate 2019) <https://www.loc.gov/law/help/fake-news/counter-fake-news.pdf>; Harriss L and Raymer K, "Online Information and Fake News" (POST, UK Parliament, July 12, 2017) <https://researchbriefings.parliament.uk/ResearchBriefing/Summary/POST-PN-0559>

(5) "Tackling the Information Crisis: A Policy Framework for Media System Resilience" (LSE Truth, Trust & Technology Commission 2018) <http://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis.pdf>; Vīķe-Freiberga V and others, "A Free and Pluralistic Media to Sustain European Democracy" (The Report of the High Level Group on Media Freedom and Pluralism 2013) <https://ec.europa.eu/information_society/media_taskforce/doc/pluralism/hlg/hlg_final_report.pdf>

(6) Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies, Para 7(a)&(b)

and ensure "greater transparency".(1) Also, all upcoming legal frameworks must be more inclusive to keep pace with the "fluid media environment" and to cover "all types of journalistic activities" on all mediums.(2) On top of that, an independent platform committee, which we suggest to be of an international character, should be established to report on self-regulatory measures adopted by SMPs, on how they respond to users complaints, and their redressal mechanisms.(3) Unlike the established Oversight Board by Facebook (now Meta),(4) which is a self-regulatory scheme, this proposed reviewing committee should function in a way similar to the Human Rights Committee ("HRC"),(5) for example. States, in this case, would accept the jurisdiction of the Committee, then promulgate laws that would obligate SMPs to cooperate with the Committee in relation to users' complaints the latter decides to investigate, and the Committee's decision - unlike the HRC - shall be binding upon SMPs.

When it comes to automated filtering mechanisms implemented by internet giants, it is crucial that forthcoming laws must oblige all online channels to be completely neutral when processing content by "altering the relevance algorithms that recommend, surface and suppress content to users", allowing users to suspend any personalisation activity either temporarily or permanently and promoting encouraging steps towards more exposure to opposing and diversified content.(6)

(1) "A Multi-Dimensional Approach to Disinformation" (The independent High level Group on fake news and online disinformation, European Commission 2018) <https://op.europa.eu/en/publication-detail/-/publication/6ef4df8b-4cea-11e8-be1d-01aa75ed71a1>; "Online Harms White Paper" (HR Government 2019) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf>

(2) Vīķe-Freiberga V and others, "A Free and Pluralistic Media to Sustain European Democracy" (The Report of the High Level Group on Media Freedom and Pluralism 2013) <https://ec.europa.eu/information_society/media_taskforce/doc/pluralism/hlg/hlg_final_report.pdf>

(3) "Online Harms White Paper" (HR Government 2019) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf>

(4) The Oversight Board helps Meta "answer some of the most difficult questions around freedom of expression online: what to take down, what to leave up and why." by "[reviewing] a selected number of highly emblematic cases and determine if [content moderation] decisions were made in accordance with Facebook›s stated values and policies." Please, visit https://oversightboard.com/ for further information

(5) ‹Individual Communications› (OHCHR) <https://www.ohchr.org/en/treaty-bodies/ccpr/individual-communications>

(6) Sunstein CR, #Republic: Divided Democracy in the Age of Social Media (Princeton University Press 2017);

Further, the adoption of "appropriate legislative, regulatory and supervisory frameworks related to algorithmic systems" is of great necessity, and if algorithmic systems or such are used by private actors, they must "comply with the applicable laws and fulfil their responsibilities to respect human rights in line with the UN Guiding Principles on Business and Human Rights and relevant regional and international standards".(1)

The existence of meaningful human intervention is of utmost importance, in other words the "human-in-command" approach is important to apply, where the human should be allowed to retain control over the systems at any given point.(2) There are not many cases concerning the amount required for human intervention yet there is the case of Muthukumar v. Telecom Regulatory Authority of India & Ors. before the Madras High Court in India.(3) It concerned banning the downloads of the Tik Tok application due to hosting inappropriate content for the culture. The content moderation on Tik Tok consisted of several layers, it starts off by using an AI filter to ensure that the content is appropriate, and then three levels of human moderation, in addition to being able to send complaints to human reviewers.(4) The court came to the conclusion that the application had sufficient human involvement due to those multiple layers, as well as using multiple languages to review the content.(5)

With this sufficient human involvement, self-regulation can be allowed.

"Tackling the Information Crisis: A Policy Framework for Media System Resilience" (LSE Truth, Trust & Technology Commission 2018) <http://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis.pdf>; Vīķe-Freiberga V and others, "A Free and Pluralistic Media to Sustain European Democracy" (The Report of the High Level Group on Media Freedom and Pluralism 2013) <https://ec.europa.eu/information_society/media_taskforce/doc/pluralism/hlg/hlg_final_report.pdf>

(1) CDMSI, Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies, para 3

(2) European Economic and Social Committee ("EESC") 526th EESC plenary session of 31 May and 1 June 2017 'Opinion of the European Economic and Social Committee on 'Artificial intelligence — The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society'' 2017/C 288/01 31 August 2017 para 1.6

(3) WP(MD) No. 7855 of 2019 (24 April 2019)

(4) Muthukumar v. Telecom Regulatory Authority of India & Ors. WP(MD) No. 7855 of 2019 (Madras High Court 24 April 2019) para 4 & 10

(5) Muthukumar v. Telecom Regulatory Authority of India & Ors. WP(MD) No. 7855 of 2019 (Madras High Court 24 April 2019) para 12

In other words, the installation of filters should only be allowed under two conditions: 1) "for the protection of minors, in particular in places accessible to them"; and/or 2) the filter should comply with the standards of freedom of expression and access to information and "measures may be taken to enforce the removal of clearly identifiable internet content or, alternatively, the blockage of access to it, if the competent national authorities have taken a provisional or final decision on its illegality."(1) Emphasis should be paid to the phrase "the competent national authorities have taken a provisional or final decision on its illegality". Thus, disallowing private control of rights.

As it stands, this should not put a monitoring obligation on SMPs, as the filters are susceptible to errors or just not being able to detect hateful content.(2) This supports the aforementioned stance of the CJEU,(3) with additional steps that are mentioned in the paragraph above, over the also aforementioned stance of the ECtHR(4) when it comes to filtering content.

iv .Conclusion

In this digital era, Social Media Platforms play a major role in shaping public discourse. Personalisation features could arguably be beneficial, nonetheless, it generates unignorable dire implications on democracy. Whilst some may claim that the ramifications of automated online content filtering are not that serious, it is crystal clear that echo-chambers and filter bubbles carry greater disadvantages, resulting in a polluting the digital sphere with discrimination, disinformation, extremism, manipulation, and political polarisation. Simply put, they are bringing people back to the time of private forums and secret chat-groups which were used to embrace societal fragmentation, help radicalise individuals, and inflame hatred.

(1) COUNCIL OF EUROPE COMMITTEE OF MINISTERS DECLARATION on freedom of communication on the Internet (Adopted by the Committee of Ministers on 28 May 2003 at the 840th meeting of the Ministers› Deputies) Principle 3

(2) Delfi AS v Estonia App no 40287/98 (ECtHR 16 June 2015) para 156

(3) Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM) Case C-70/10 (CJEU 24 November 2011) Para 53; Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV Case C-360/10 (CJEU 16 February 2012) para 52

(4) Delfi AS v Estonia App no 40287/98 (ECtHR 16 June 2015) para 156

Also, upload filters contribute to all the implications of automated filtering by propagating permanent, automated censorship. All this has eventually led to a novel form of censorship namely "surveillance capitalism" which was first coined by Professor Shoshana Zuboff when she offered a disturbing picture of how internet giants exploit users' personal data as free raw material "not only to predict [users] behaviour but also to influence and modify it; and how this has had disastrous consequences for democracy and freedom".(1)

It is imperative to stress the fact that the right to receive and impart information is an integral part of the right to freedom of expression. Thus, the deployment of automated online filtering, while may not directly go against every human rights treaty, yet in a more elaborate manner, goes against the principles of freedom of expression. That could be witnessed by consulting human rights instruments, and in a more direct sense, Article 13 of the ACHR which clearly demonstrates that completely automated filtering violates freedom of expression and all other rights derived from it. This also could be inferred through the examination of a variety of soft law documents.(2)

Correspondingly, the international community must, on the one hand, adopt substantial regulatory measures, emanating from international human rights law, aimed at alleviating these serious implications and actively take effective and immediate steps to push internet giants to adopt policies that would remedy the existing dilemma for a better online atmosphere. States and Digital Rights Organisations should push against the automatic activation

(1) Joanna Kavenna, 'Shoshana Zuboff: Surveillance capitalism is an assault on human autonomy' (The Guardian 2019) <https://www.theguardian.com/books/2019/oct/04/shoshana-zuboff-surveillance-capitalism-assault-human-automomy-digital-privacy>; For further information, please, consult Donell Holloway, 'Explainer: what is surveillance capitalism and how does it shape our economy?' (The Conversation 2019) <https://theconversation.com/explainer-what-is-surveillance-capitalism-and-how-does-it-shape-our-economy-119158>

(2) "OAS: JOINT DECLARATION ON FREEDOM OF EXPRESSION AND THE INTERNET" (Oas.org, 2011) Para 1 (B) <http://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1>; "Unboxing artificial intelligence: 10 steps to protect human rights" (Council of Europe Commissioner for Human Rights 2019) P.19; Muthukumar v. Telecom Regulatory Authority of India & Ors WP(MD) No. 7855 of 2019 (Madras High Court 24 April 2019); COUNCIL OF EUROPE COMMITTEE OF MINISTERS DECLARATION on freedom of communication on the Internet (Adopted by the Committee of Ministers on 28 May 2003 at the 840th meeting of the Ministers› Deputies) Principle 3; See e.g. Francisco Martorell v. Chile Case 11.230 Report No. 11/96 Inter-Am.C.H.R.,OEA/Ser.L/V/II.95 Doc. 7 rev. at 234 (IACHR 1997) para. 58&59; Steve Clark v. Grenada Case 10.325 Report No. 2/96 Inter-Am.C.H.R., OEA/Ser.L/V/II.91 Doc. 7 at 113 (IACHR 1996) para 8

of algorithmic content personalisation, demanding that all SMPs offer their users the option to opt-in to or opt-out of automated content personalisation.(1) One the other hand, educational campaigns with a focus on media literacy and diversity studies are a vital key player in combating disinformation, censorship, political isolation, and social extremism.

Furthermore, great attention should be paid to the current regulatory tensions evidenced by the three-sided encounter between tech companies, the international community, and national and legislative authorities with each trying to claim jurisdictional authority over online content regulation. That is because while internet giants frequently claim that human rights standards are embedded in their policies and that their activities are deployed after thorough human rights scrutiny, they have attempted on different occasions to avoid international, regional, and national regulatory frameworks by pushing for a self-regulatory scheme. As previously mentioned, a clear example would be Meta's Oversight Board.

Finally, and as a response to such self-regulatory moves, it is pivotal that there should be a collective agreement on employing international human rights law as the bedrock for any States-sponsored regulatory framework. This is instead of attempting to take individual actions that would result in jurisdictional differences, making it unattainable for tech companies to conform to.

(1) Donahoe E and Metzger M, 'Artificial Intelligence and Human Rights' (2019) 30 Journal of Democracy 124; Harambam J, Bountouridis D, Makhortykh M and Van Hoboken J, 'Designing for the better by taking users into account: a qualitative evaluation of user control mechanisms in (news) recommender systems', In Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19) (Association for Computing Machinery, New York, USA, 2019) 69-77; Oremus W, 'Twitter Has Finally Made It Easy to Set Your Timeline to Reverse-Chronological' (Slate, 2018) <https://slate.com/technology/2018/12/twitterreverse-chronological-timeline-setting.html>