FILOSOFIJA. SOCIOLOGIJA
2024. T. 35. Nr. 1, p. 67–74

ISSN 0235-7186 eISSN 2424-4546
DOI: https://doi.org/10.6001/fil-soc.2024.35.1.9

# New (Digital) Media in Creative Society: Ethical Issues of Content Moderation

SALVATORE SCHINELLO
Department of Philosophy and Cultural Studies, Faculty of Creative Industries, Vilnius Gediminas Technical University, 1 Trakų Street, 00132 Vilnius, Lithuania
Email: salvatore.schinello@vilniustech.lt

Digitalisation and platformisation are continuously impacting and reshaping the societies we live in. In this context, we are witnessing the rise of phenomena such as fake news, hate speech, and the sharing of any other illegal content through social media. In this paper, I propose some ethical reflections on content moderation in the context of digital (social) media, as this topic seems – to me – to already incorporate other relevant digital issues in it, such as algorithms bias, the spread of fake news, and the potential misuses of artificial intelligence. In the first section, I will provide a few hermeneutic reflections over a speech given by the Italian scholar Umberto Eco, which appears to underline the necessity of a content moderation in an era of digital (social) media. In the second section, I will analyse, through a consequentialist perspective, critical and ethical issues posed by content moderation. In particular, I suggest the idea of a 'moderate' (reasonable and limited) content moderation that can only be assured by humans, as they are able to contextualise the content, to take emotions and subjective elements into account, to apply critical thinking and adaptability in complex circumstances.

**Keywords:** digital media, social media, content moderation, ethical issues and challenges, Umberto Eco, censorship

## INTRODUCTION

New media and technologies play a relevant role in today's (creative) societies (Kačerauskas 2015). We are witnessing what philosopher V. Flusser, almost prophetically, referred to as a 'telematic revolution' produced by the interconnection of telecommunications and informatics (Flusser 2011). On the grounds of this, Flusser anticipated the dominance of the digital image over the text and a shift in the perception of reality (the phenomena of echo chambers and filter bubbles) in social media. Similarly, a profound change in the passage from 20th century's mass media to digital media is observed by J. D. Peters (2015), as the main innovation brought by the latests does not consist in transmitting contents but 'in tracking, tweeting, and tagging, in the structures of everyday life and the organization of power' (Peters 2015: 7). This political aspect of digital media may, in a certain way, be also connected to the critical tradition of communication developed by R. T. Craig (1999) in his seminal work *Communication Theory as a Field*.

The advent of digital and, among them, social media – on the one hand – and the increasing employment of platforms in various aspects of our everyday life – on the other – have created two phenomena that are continuously impacting and reshaping the societies we live in: digitalisation and platformisation (Magaudda, Solaroli 2021). Both of them are by some means interconnected in the effort of leading today's society through a digital transformation and bring a set of ethical issues and challenges that must be taken in careful consideration by scholars, governments and businesses.

Among all the ethical issues that can be related to the progressive digital transformation of society we can certainly mention the accumulation and the protection of personal data in the context of what scholars consider to be a 'surveillance capitalism' (Andrew, Baker 2021), cybersecurity threats (Snider et al. 2021), unequal access to digital technologies producing economic disparities and new social exclusions (Lythreatis et al. 2022), the impact of automation on employment (Filippi et al. 2023), health issues such as digital addiction and sleep disorders (Dresp-Langley, Hutt 2022), and the phenomenon of the e-waste (Patil, Ramakrishna 2020).

In this paper, my aim is to propose some ethical reflections on content moderation in the context of digital (social) media, as this topic seems – to me – to already incorporate other relevant digital issues in it, such as algorithms bias, the spread of fake news and the potential misuses of artificial intelligence.

In the first section, I will provide a few hermeneutic reflections over a speech given by the Italian scholar Umberto Eco, which appears to underline the necessity of a content moderation in an era of digital (social) media. In the second section, I will apply a consequentialist approach, which takes into account the results and consequences of any action when evaluating its rightness or wrongness (Shaw 2006), to the analysis of critical and ethical issues posed by content moderation.

## THE NEED FOR A CONTENT MODERATION IN AN ERA OF DIGITAL MEDIA

Why is a content moderation necessary for a healthy and safe digital environment? During his speech at the ceremony for the conferment of a Doctorate 'Honoris Causa' in Communication and Culture from the University of Turin in 2015, the Italian scholar Umberto Eco, using his well-known biting and ironical style, affirmed: 'Social networks give legions of idiots the right to speak when they once only spoke at a bar after a glass of wine, without harming the community. Then they were quickly silenced, but now they have the same right to speak as a Nobel Prize winner' (Kristo 2017: 52). My suggestion here is to go beyond Eco's provocative and harsh tone and to focus on a few significant elements that emerge from this extract of his speech.

First of all, he affirms that, due to social networks, the right of speech has been given to 'legions of idiots'. We certainly understand that, beyond the provocation and having been him in life a sincere defender of democratic values, Eco is not putting in discussion the freedom of speech, which is a fundamental right that belongs not only to Nobel Prize winners but to all human beings (even to those he refers to as 'idiots'), and is solemnly recognised by the Universal Declaration of Human Rights. What he seems to be questioning is, instead, the facility, velocity and gratuitousness with which a content that is potentially harmful to the community is spread, as a virus, thanks to social media. In this case, we speak, not inappropriately, of 'virality' of content (Mills 2012).

The second element I want to focus on is the image of the 'glass of wine' that would encourage the so-called 'idiots' to express themselves. On the one hand, we can remember the Latin phrase '*In vino veritas*', which suggests the idea that, if inebriated by the wine, the actors of a communicative interaction are open, frank and sincere to each other. On the other hand, the lowering of filters and the loosening of inhibitions prevent people from having a full and sober control over what they are communicating (by saying, writing or sharing through social media).

In the environment of digital (social) media, the encouraging – and, in this case, negative – effect of the 'glass of wine' is assured by the screen of our electronic devices that, firstly, by dehumanising the other (who does not physically stand in front of us and is perhaps unknown to us) and, secondly, by de-individuating us (through fake profiles that guarantee anonymity), gives us the impression of being authorised to express the worst of ourselves without any responsibility, producing a sort of 'Lucifer Effect' (Zimbardo 2007).

Only moderation (of tones and content) can re-establish the common sense and guarantee a sober and health ('*In aqua sanitas*') communication environment.

In his aforementioned speech, U. Eco speaks of a content that has the potential to be harmful not only to the communicative interaction itself but to the community as a whole. Here we can merely mention two phenomena that gained momentum due to the diffusion of digital (social) media: misinformation (the spread of fake news) and hate speech (Cinelli et al. 2021).

Finally, this extract of Eco's speech appears to underline the necessity of silencing (moderating) a content that is potentially harmful to the community, and to demand for a content moderation, in an era of digital (social) media, as a tool which ensures that the content being shared is not dangerous, inappropriate or illegal.

In doing so, content moderation should ideally contribute to strengthen the stability of democracy. As Eco himself states: '<...> democracy also means accepting a tolerable quantity of injustice to avoid greater injustice' (Eco 2014: 101). Content moderation is perhaps the tolerable injustice that today's society can endure to prevent more serious injustice, such as the diffusion of false information and the incitement of violence and hate. However, content moderation itself requires 'moderation' so as not to become a mere censorship and a digital authoritarianism.

In the next section, I will analyse critical and ethical issues posed by content moderation.

## FROM CONTENT MODERATION TO CONTENT CENSORSHIP?

Despite the undoubtedly fundamental importance of content moderation in making the digital environment a safe and health place, we cannot fail to recognise that this process can lead to forms of authoritarianism and arbitrariness.

First of all, we should analyse the ways in which content moderation can be conducted. Generally speaking, we can distinguish between manual (human), automated, and hybrid content moderations (Wang, Kim 2023).

The automated content moderation is characterised by the use of algorithms and artificial intelligence in order to automatically identify and filter the content that conflicts with the community's guidelines.

It is difficult to quantify the amount of user generated content (USG) that is posted on various social media every day; however, we can easily imagine that it deals with large quantities that cannot be managed by human beings alone. As a result, the artificial intelligence is progressively integrated in content detection and moderation, giving birth to automated or hybrid (AI/human) strategies.

All social media platforms are putting efforts in moderating what is considered to be an unsuitable content that violates community's guidelines or even the law. Platforms should be able to identify and, subsequently, moderate the inappropriate content. Once the latter is detected (through human, AI or collaborative detection) the actions that social media can take are the following: social media may remove the content or, for more serious cases, they can also decide to block for a certain period of time, or definitely remove, the account through which the content was published (Gongane et al. 2022).

We can suppose that content moderation, especially in non-democratic regimes, could be used as an instrument of digital repression. Scholars Luo and Li (2022) study the environment of social media in China and their contribution in supporting the authoritarian rule by encouraging peer-censorship among online communities' members. In fact, digital censorship does not only operate from top to bottom but, by involving the users themselves in accusatory reporting, is becoming a 'collective work' (Luo, Li 2022: 3) which integrates state censorship in controlling the spread of alternative opinions and information. Authors explain this mechanism as follows: '[d]ifferent factions of the community strategically borrowed language and practices from the political authority to discipline content they disliked, and every member could simultaneously be the subject and the object of accusatory reporting' (Luo, Li 2022: 15).

Digital censorship phenomena are not just a prerogative of authoritarian regimes; they can also be observed in liberal (western) democracies as a consequence of new strict legislations whose declared objective is to build a safer digital environment through a deep content moderation.

In the context of the European Union, for instance, there is a call for a transition from the current self-regulation (where social media platforms autonomously define the guidelines for content moderation) to a transnational political regulation (whose guidelines derive from the EU legislation). This is the case of the so-called Digital Services Act, a EU Regulation that will become effective in 2024. According to Schlag (2023), through this new regulation of social media platforms, the EU aims at reestablishing its digital sovereignty and 'taking back control from big and US-based enterprises' (Schlag 2023: 169). This may be interpreted as a political interference in the public sphere (which is, by definition, independent of the political control and open to all members of society); however, platforms themselves 'are not truly independent of government control and corporate influence, as they are privately owned and can be subject to censorship and manipulation' (Schlag 2023: 174). Nevertheless, every new (public or private) regulation of the digital environment, as it may contract the spaces of freedom, demands to be taken in serious consideration.

Remaining in the context of Western democracies, forms of self-censorship are observed by scholars in Canada and North America (Hu, Barradas 2023). Self-censorship is defined as 'the act of limiting or controlling one's own expression or behaviour to avoid offending or upsetting others, or to conform to social or cultural norms' (Hu, Barradas 2023: 609). If we perceive it as a form of self-limitation (moderation) when it comes to protect the most vulnerable segments of society, it cannot but be warmly accepted. Nevertheless, the aforementioned definition of self-censorship explains that this phenomenon can also be understood as a form of conformism to hegemonic social or cultural norms and values. The question arises: Are social media users self-censoring themselves because they are conscious of the necessity of behaving well and respectfully on social media, or because they are simply afraid of being

banned, of becoming object of any other form of soft moderation (such as warning labels on their posts) if they share a content that is considered unpopular or 'controversial', or even of having disciplinary repercussions in their workplace?

In the context of crises, such as wars or sociopolitical conflicts, social media 'perform a critical civic role as spaces for eyewitness testimony, news and information, humanitarian efforts, collective action, and legal accountability' (Lewis 2023: 2398). In other words, in today's society social media play the fundamental role of integrating traditional media in the effort of bringing to light facts and events that would probably be undivulged, and, in doing so, they make the public opinion aware on what is happening in the world. Nevertheless, when it comes to share information that is related to war or social disorders – especially if in photos or videos – we are faced with a content that is likely to contain violence, and that can potentially be recognised as dangerous by the algorithms on which automated content moderation is based, and, consequently, censored.

This indicates that social media platforms possess, at this point, the power to determine 'what counts as fact and truth <…> but also reveals the ways platforms make value judgments and moral choices' (Lewis 2023: 2413). In fact, the algorithms underlying automated content moderation can hardly differentiate the denunciation of violence from the promotion of violence, or the artistic nudity from the explicit sexual content; through their act of content moderation they delineate the boundaries of what can be considered moral, tolerable or acceptable in the digital space.

The issues of content moderation are closely related to the monitoring and prevention of the so-called fake news. Stewart (2021) affirms the following:

> 'Accurately identifying fake news requires that interested parties agree on what makes fake content problematic, such that it merits removal, and that content moderators can reliably distinguish this content from non-problematic content' (Stewart 2021: 923).

In other words, both users and platforms should have the same perception over what is 'fake' and over the necessity of removing a content that – being untrue and malicious – contributes to the spread of misinformation across platforms and, consequently, societies. However, only achieving equilibrium between guaranteeing users the right to freely and creatively express their own opinions, and cancelling only the content that is truly harmful (a reasonable, limited and therefore 'moderate' content moderation) can social media maintain the trust and loyalty of their users. Otherwise, the latter could feel that their freedom is limited by a technology that is only formally 'open' and can decide to leave in favour of a different competing platform.

Yet, I believe that a true 'moderate' content moderation can only be assured by humans, as they are able to contextualise the content, to take emotions and subjective elements into account, to apply critical thinking and adaptability in complex circumstances.

Analysing the case of unjustified account deletions on social media platforms Instagram and TikTok, researcher C. Are (2023) identifies three possible solutions that might prevent from content moderation abuses:

> (1) *Improved transparency*. Users should be fully informed about the communities' guidelines and the motivations behind the decision taken by the platforms to delete or shadowban their accounts. This implies a more individualised communication towards users, which, however, might be financially onerous for the platforms.

(2) *Recognition of malicious flagging as a form of online abuse.* This solution presents, in my humble opinion, some criticisms, as it is not easy to establish whether a report is malicious or not. In other words, every report arises from the fact that a certain user feels damaged or annoyed by a certain content, and all of this is purely subjective. Nevertheless, platforms must ensure that the reported content (or the account) will be restored if after analysing the report the content turns out to comply with the guidelines.

(3) *Investment in 'deleted creators' communication teams.* The aim of this process would be to improve the user support and assistance, and contribute to a greater transparency towards users. However, a question arises: Is it preferable to invest in user support or in manual (human) content moderation?

After having analysed some criticisms presented by content moderation, I would like to conclude on the impact that this phenomenon can have on societies that are considered to be 'creative'. If we agree that one of the characteristics of the creative society is its openness (Kačerauskas 2017), than a massive and automated content moderation, operated by algorithms, can be a threat to it. As we noticed in this part of the paper, content moderation presents criticisms such as peer-censorship and self-censorship phenomena, and unjustified content removal: all of them can prevent creators from publishing a creative content that may be potentially considered controversial or provocative, limiting *de facto* their artistic and civic freedom.

To be functional to the creative society, the environment of digital (social) media shall preserve its openness and its tolerance to contents that can be, in a sense, susceptible to questioning existing norms and values, without having to give up its objective of protecting the most vulnerable segments of society and making the digital environment a safe and health place for all users.

## CONCLUSIONS

Due to the processes of platformisation and digitalisation, digital (social) media have become fundamental communication tools, used by billions of users worldwide. Nevertheless, we are witnessing the rise of phenomena such as fake news, hate speech, and the sharing of any other illegal content through social media. If it is true, as U. Eco affirms, that '[s]ocial networks give legions of idiots the right to speak' (Kristo 2017: 52), then the need for a content moderation comes from this.

Content moderation can be performed by humans, machines (AI algorithms) or by a combined action of both humans and machines. On the one hand, manual (human) content moderation is financially onerous to platforms; on the other hand, automated content moderation can potentially lead to censorship. Peer-censorship, self-censorship and unjustified content removal combined with an excessive (both public and private) regulation of this sector, can make the digital environment a space where every user is formally free to express him/herself only to the extent that the algorithms underlying automated content moderation do not label it as harmful content. This may prevent users from sharing denouncing content (that can contain elements of violence in the context of war or social disorders) or creative content (that can be provocative or controversial and, for this reason, removed).

We can recognise, using Eco's words, that content moderation is 'a tolerable quantity of injustice to avoid greater injustice' (Eco 2014: 101), such as misinformation, hate speech, cyberbullying and the promotion of violence. Nonetheless, a massive automated content

moderation can give birth to a toxic digital environment based on censorship and manipulation, and this may represent a threat to the creative society, which is, by definition, open and tolerant.

In this paper, I suggest the idea of a 'moderate' content moderation that, in appropriate conditions, can only be assured by humans, as they are able to contextualise the content, to take emotions and subjective elements into account, to apply critical thinking and adaptability in complex circumstances. Implementing such a content moderation would require the platforms to make a financial effort in order to increase their human resources; however, this seems to be an effective means of ensuring a balance between the freedom of expression and the necessity to maintain a safe digital space.

Nevertheless, there are still some questions remaining. Are social media platforms available to increase their investments in human resources? Is this financially sustainable for them? How are content moderators selected and hired? Are human content moderators able to go beyond their subjective judgments and guarantee an objective moderation?

Received 29 November 2023
Accepted 18 January 2024

## References

1. Andrew, J.; Baker, M. 2021. 'The General Data Protection Regulation in the Age of Surveillance Capitalism', *Journal of Business Ethics* 168(3): 565–578. Available at: https://doi.org/10.1007/s10551-019-04239-z
2. Are, C. 2023. 'An Autoethnography of Automated Powerlessness: Lacking Platform Affordances in Instagram and TikTok Account Deletions', *Media, Culture & Society* 45(4): 822–840. Available at: https://doi.org/10.1177/01634437221140531
3. Cinelli, M.; Pelicon, A.; Mozetič, I.; Quattrociocchi, W.; Novak, P. K.; Zollo, F. 2021. 'Dynamics of Online Hate and Misinformation', *Scientific Reports* 11: 22083. Available at: https://doi.org/10.1038/s41598-021-01487-w
4. Craig, R. T. 1999. 'Communication Theory as a Field', *Communication Theory* 9(2): 119–161. Available at: https://doi.org/10.1111/j.1468-2885.1999.tb00355.x
5. Dresp-Langley, B.; Hutt, A. 2022. 'Digital Addiction and Sleep', *International Journal of Environmental Research and Public Health* 19(11): 6910. Available at: https://doi.org/10.3390/ijerph19116910
6. Eco, U. 2014. *Turning Back the Clock: Hot Wars and Media Populism*. New York: Random House.
7. Filippi, E.; Bannò, M.; Trento, S. 2023. 'Automation Technologies and Their Impact on Employment: A Review, Synthesis and Future Research Agenda', *Technological Forecasting and Social Change* 191: 122448. Available at: https://doi.org/10.1016/j.techfore.2023.122448
8. Flusser, V. 2011. *Into the Universe of Technical Images*. Minneapolis: University of Minnesota Press.
9. Gongane, U. V.; Munot, M. V.; Anuse, A. D. 2022. 'Detection and Moderation of Detrimental Content on Social Media Platforms: Current Status and Future Directions', *Social Network Analysis and Mining* 12: 129. Available at: https://doi.org/10.1007/s13278-022-00951-3
10. Hu, W.; Barradas, D. 2023. 'Work in Progress: A Glance at Social Media Self-Censorship in North America', in *Proceedings of the 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 609–618. Available at: https://doi.org/10.1109/EuroSPW59978.2023.00072
11. Lewis, K. 2023. 'Colonising the Narrative Space: Unliveable Lives, Unseeable Struggles and the Necropolitical Governance of Digital Population', *Information, Communication & Society* 26(12): 2398–2418. Available at: https://doi.org/10.1080/1369118X.2023.2230260
12. Luo, Z; Li, M. 2022. 'Participatory Censorship: How Online Fandom Community Facilitates Authoritarian Rule', *New Media & Society*. Available at: https://doi.org/10.1177/14614448221113923
13. Lythreatis, S.; Singh, S. K.; El-Kassar, A.-N. 2022. 'The Digital Divide: A Review and Future Research Agenda', *Technological Forecasting and Social Change* 175: 121359. Available at: https://doi.org/10.1016/j.techfore.2021.121359
14. Kačerauskas, T. 2017. *Kūrybos visuomenė*. Vilnius: Technika.
15. Kačerauskas, T. 2015. 'Technologies in Creative Economy and Creative Society', *Technological and Economic Development of Economy* 21(6): 855–868. Available at: https://doi.org/10.3846/20294913.2015.1036325

16. Kristo, R. M. 2017. 'Umberto Eco and Emotions in the Time of Internet', *International Journal of Social and Educational Innovation* 4(7): 51–58.
17. Magaudda, P; Solaroli, M. 2021. 'Platform Studies and Digital Cultural Industries', *Sociologica* 14(3): 267–293. Available at: https://doi.org/10.6092/issn.1971-8853/11957
18. Mills, A. J. 2012. 'Virality in Social Media: The SPIN Framework', *Journal of Public Affairs* 12(2): 162–169. Available at: https://doi.org/10.1002/pa.1418
19. Patil, R. A.; Ramakrishna, S. 2020. 'A Comprehensive Analysis of e-Waste Legislation Worldwide', *Environmental Science and Pollution Research* 27: 14412–14431. Available at: https://doi.org/10.1007/s11356-020-07992-1
20. Peters, J. D. 2015. *The Marvelous Clouds: Towards a Philosophy of Elemental Media*. Chicago: University of Chicago Press.
21. Schlag, G. 2023. 'European Union's Regulating of Social Media: A Discourse Analysis of the Digital Services Act', *Politics and Governance* 11(3): 168–177. Available at: https://doi.org/10.17645/pag.v11i3.6735
22. Shaw, W. 2006. 'The Consequentialist Perspective', in *Contemporary Debates in Moral Theory*, ed. J. L. Dreier. Oxford: Blackwell, 5–20.
23. Snider, K. L. G.; Shandler, R.; Zandani, S.; Canetti, D. 2021. 'Cyberattacks, Cyber Threats, and Attitudes Toward Cybersecurity Policies', *Journal of Cybersecurity* 7(1): tyab019. Available at: https://doi.org/10.1093/cybsec/tyab019
24. Stewart, E. 2021. 'Detecting Fake News: Two Problems for Content Moderation', *Philosophy & Technology* 34: 923–940. Available at: https://doi.org/10.1007/s13347-021-00442-x
25. Wang, S.; Kim, K. J. 2023. 'Content Moderation on Social Media: Does it Matter Who and Why Moderates Hate Speech?', *Cyberpsychology, Behavior, and Social Networking* 26(7): 527–534. Available at: https://doi.org/10.1089/cyber.2022.0158
26. Zimbardo, P. 2007. *The Lucifer Effect: Understanding How Good People Turn Evil*. New York: Random House.

SALVATORE SCHINELLO

# Naujosios (skaitmeninės) medijos kūrybos visuomenėje: turinio moderavimo etiniai klausimai

*Santrauka*

Skaitmenizacija ir platformizacija nuolat paveikia ir pertvarko visuomenę, kurioje gyvename. Kartu pastebime tokių reiškinių, kaip melagingų naujienų, neapykantą kurstančių kalbų ir bet kokio kito neteisėto turinio dalijimosi augimą socialiniuose tinkluose. Šiame straipsnyje siūlomi keletas etinių apmąstymų, susijusių su turinio moderavimu skaitmeninių (socialinių) medijų kontekste, nes ši tema apima kitus svarbius skaitmenizacijos klausimus, kaip antai algoritmų šališkumas, melagingų naujienų plitimas ir galimai netinkamas dirbtinio intelekto naudojimas. Pirmoje straipsnio dalyje pateikiamas hermeneutinis apmąstymas apie italų mokslininko Umberto Eco kalbą, kuri pabrėžia turinio moderavimo būtinybę skaitmeninių (socialinių) medijų epochoje. Antroje dalyje nagrinėjamos kritinės ir etinės turinio moderavimo problemos iš konsekvencialistinės perspektyvos. Siūloma turinio „moderuoto" (nuosaikaus ir riboto) moderavimo idėja: šis moderavimas galėtų būti užtikrintas žmonių, kurie, kitaip nei mašinos, sugeba kontekstualizuoti turinį, atsižvelgti į emocijas ir subjektyvius elementus, kritiškai mąstyti ir prisitaikyti esant sudėtingoms aplinkybėms.

**Raktažodžiai:** skaitmeninės medijos, socialiniai tinklai, turinio moderavimas, etiniai klausimai ir iššūkiai, Umberto Eco, cenzūra