

Content moderation in (decentralised) metaverses

**Proceedings of the International Congress Towards a Responsible
Development of the Metaverse, 13-14 June 2024, Alicante**

Gian Marco Bovenzi

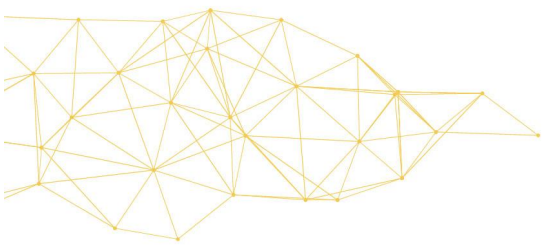
Centro Alti Studi per la Difesa, Rome



The Chair for the Responsible Development of the Metaverse (MetaverseUA Chair) was created by the University of Alicante (Spain) and financed by Meta Platforms under its [XR Program and Research Funds](#). The Program aims at supporting academic and independent research across Europe into metaverse issues and opportunities. The MetaverseUA Chair is a member of the [European Metaverse Research Network](#). Like all our work, this report has been produced completely independently. The ideas expressed in this paper are the sole responsibility of the author.

How to cite this paper:

Bovenzi, G.M., 'Content moderation in (decentralized) metaverses.' (2024) *Proceedings of the International Congress Towards a Responsible Development of the Metaverse*, Alicante, 13-14 June, 2024.



Abstract

In implying the obligation for hosting service providers, upon obtaining actual knowledge of illegal activities or content shared on an online platform, to remove such content, the current enforced regulations on content moderation are intended to provide safer online environments, contextually protecting free speech and other fundamental rights of their users. Despite the significant regulatory steps taken so far, it is likely that the metaverse and virtual environments might raise further issues in the field, both with regards to the typologies of illegal content shared on a platform and to the actual possibility of removing a given illegal or harmful content in decentralized platforms, thus entailing enhanced risks for users. Given the novelty, the absence of enforced regulations, and the scarce literature on the topic, this article provides a review of the existing literature on content moderation in decentralized platforms, aiming at contributing to the debate and offering discussion on policy solutions.

Keywords: Content moderation, Digital Services Act, metaverse, decentralized platforms, decentralized social media, decentralized metaverses

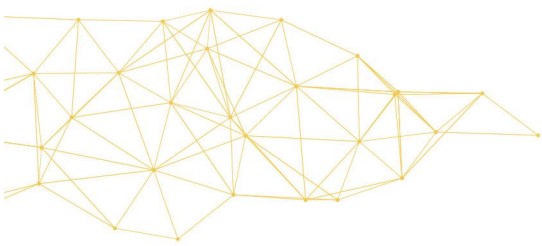
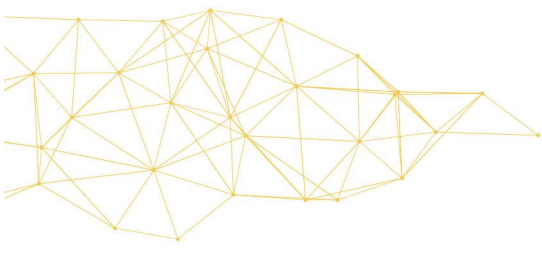


Table of Contents

1. Introduction	1
2. Content and conduct moderation in the EU framework	3
3. The metaverse, the decentralized web and the issues for content moderation	5
3.1. Revisiting content moderation in light of virtual environments	8
3.2. Content moderation in decentralized platforms	10
4. Content moderation in decentralized metaverses	13
5. Conclusions	14



1. Introduction

Content moderation on the internet and social media platforms has been broadly defined as the ‘governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse’,¹ thus implying the deep control of user-generated contents (UGC) potentially resulting in illegal or harmful carried out by service providers and internet intermediaries. A content shared by a user on a digital platform is evaluated, upon its notice, by the intermediary services providers, owners of the platform on which the content is shared, and it is considered inappropriate either when incompatible with the platform’s terms and conditions, or when the content violates general legal provisions². More precisely, content moderation can then be defined as ‘the screening, evaluation, categorization, approval or removal [...] of online content according to relevant communications and publishing policies’³.

The ultimate rationale of content moderation is to avoid that illegal contents might result harmful for the online community and, thus, the practice of moderation results in making such content unavailable, inaccessible, or eventually removed from the platform in order not to be visualized by platform users. Typologies of illegal contents might include texts (libel, threat, obscene content, harassment or discrimination), images or videos (pornography, assault, violent conducts), audio (mostly when it comes to manipulated audios infringing copyrights), but also online incitement to crime or terrorist propaganda and organization. The risk of illegal content is perhaps enhanced by the diffusion of AI-tools – let us think, for instance, about deep-fakes – capable of diversifying its typologies.

Nevertheless, sharing online content represents a manifestation of the right to free speech and platform users’ fundamental rights and, as such, the governance of content moderation needs to provide mechanisms that ensure a fair and transparent moderation process, limiting censorship and possible arbitrary scrutiny carried out by the internet intermediaries: therefore, services’ policies on content moderation should include ‘clear and specific statements of reasons for their content moderation decisions’⁴. The central role of platforms’ users in content sharing is also ensured via democratic moderation processes, that foresee their active participation in reporting illegal or harmful content⁵

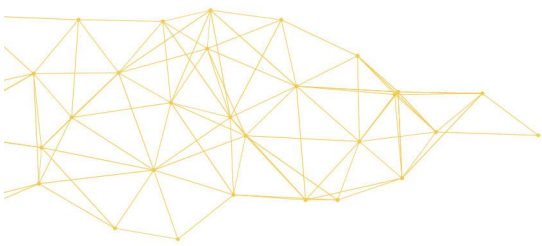
¹ J. Grimmelmann, ‘The Virtues of Moderation’ (2015) 17 *Yale Journal of Law & Technology* 42

² Absent being worldwide uniform rules of what can be considered illegal, in the present paper the terms ‘inappropriate’ or ‘incompatible’ are used and, when referring to general legal provisions, it is meant to stress how a content that violates the legal framework of a given country is to be considered illegal in that country – but, perhaps, not in another country.

³ T. Flew, F. Martin and N. Suzor, ‘Internet regulation as media policy: rethinking the question of digital communication platform governance’ (2019), 10(1) *Journal of Digital Media & Policy* 33, 50.

⁴ European Commission, ‘The impact of the Digital Services Act on digital platforms’ (*European Union*, 30 April 2024) <<https://digital-strategy.ec.europa.eu/en/policies/dsa-impact-platforms>> accessed 21 May 2024

⁵ Clarification on the use of the terms ‘illegal’ and ‘harmful’ in the present paper is required and, therefore, a short clarification of a still ongoing complex debate follows: since 1996 (in the Communication of the European Communities COM(96) 487/final, issued at the very beginning of the internet era) the differentiation between illegal content and harmful content has been considered pivotal, since ‘different categories of content pose radically different issues of principle, and call for very different legal and technological responses’. Thus, a content is illegal when it violates a country’s criminal provisions, while its level of harmfulness depends much more on cultural variables. The debate on the distinction between illegal and harmful content has been going on since then, and the recent analysis carried out by the European Parliamentary Research Service (PE 649.404 of May 2020) defines illegal content as ‘a large variety of content categories that are not compliant with EU and national legislation’ while ‘potentially ‘harmful content’ refers to content which often does not strictly fall under the prohibition of a law, but might nevertheless have harmful effects’, including nudity, bullying, mis- and dis-information, fake news, and so on. In short, while illegal content falls within the provider’s removal/moderation obligations, harmful content appears to fall outside of such obligations, in being ‘information that may be



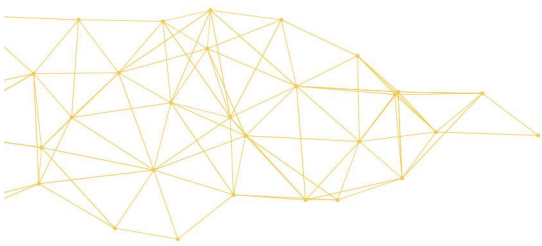
via flagging systems – starting from which the service provider, after further screening and evaluation, might eventually remove the flagged content.

In the European Union, the need for efficient regulations on content removal, balancing the right to free speech with safety and security threats posed by illegal contents, has been reflected in the recent adoption of the European Union Digital Services Act (DSA). While it is true that the DSA surely represents a pivotal tool in ensuring a safer internet, the advent of a relatively new (and relatively non-regulated) typology of digital platform, the metaverse, represents a possible game-changer and entails enhanced risks for its users in the terms that follow: first, a virtual environment could possibly reshape the concept of illegal or harmful contents itself, where the real-life likelihood of immersive experiences could dramatically increase the typologies of unwanted conducts – for instance, cyber-bullying might not only be considered as such when committed via offensive or hateful comments or speech, but also via avatar-to-avatar virtual offenses (here, the immersive environment is capable of providing further forms of psychological harm to the victim when compared to bi-dimensional content); the second arising issue concerns the decentralized internet and social networks (and thus, metaverses), where the absence of a central node controlling the network implies an actual complexity, not to say practical impossibility⁶, to remove content from a decentralized platform.

Given the novelty of the issue, this article provides an overview of the normative state of the art in the field of content moderation, focusing on the new threats posed by content sharing in the metaverse and in decentralized social media. After the introduction, the second paragraph briefly assesses the current applicable legislation in the field of content moderation in the EU framework; the third paragraph defines and describes the technical features of the metaverse and the decentralized web, highlighting the potential issues arising in the field of content moderation as better described in sub-paragraph 3.1, that offers an overview of the new typologies of illegal and harmful conducts potentially exploitable in the metaverse; and in sub-paragraph 3.2, that offers a literature review on the issues related to removing content in decentralized web architecture; finally, paragraph four wraps up the issues arising from decentralized metaverses. Given the

inadequate for certain categories of users, but whose legality varies significantly across Member States' (Francesco Vogelezang, 'Illegal vs Harmful Online Content. Some reflections on the upcoming Digital Services Act package' (*Institute for Internet and the Just Society*, 2 December 2020) <<https://www.internetjustsociety.org/illegal-vs-harmful-online-content>> accessed 28 June 2024) and thus it represents 'a delicate area with severe implication for the protection of freedom of expression' whose obligations of content removal/moderation are 'left to the discretion of intermediary service providers' (Britt van den Branden, Sophie Davidse, Eva Smit, 'In between illegal and harmful: a look at the Community Guidelines and Terms of Use of online platforms in the light of the DSA proposal and the fundamental right to freedom of expression' (*DSA Observatory*, 2 August 2021) <<https://dsa-observatory.eu/2021/08/02/in-between-illegal-and-harmful-a-look-at-the-community-guidelines-and-terms-of-use-of-online-platforms-in-the-light-of-the-dsa-proposal-and-the-fundamental-right-to-freedom-of-expression-part-1-of-3/>> accessed 28 June 2024. While this is an actual provision of the Digital Services Act, it can be stated also in relation to other regulatory frameworks). Nevertheless, this paper does not focus on the distinction between illegal and harmful content and on when providers are urged to remove a given content: therefore, for its purposes the terms 'illegal' and 'harmful' are used in a more general way, as to define/represent a content that may, in general, be harmful for the online community and thus removed from online platforms – generally in line with the UK Online Safety Act of 2023, that blurs the line between illegal and harmful contents.

⁶ B. Clifford, 'Moderating Extremism: The State of Online Terrorist Content Removal Policy in the United States' (2021), Program on Extremism, George Washington University, <<https://extremism.gwu.edu/sites/g/files/zaxdzs5746/files/Moderating%20Extremism%20The%20State%20of%20Online%20Terrorist%20Content%20Removal%20Policy%20in%20the%20United%20States.pdf>> accessed 27 April 2024.



different typologies of illegal and harmful conducts in the metaverse and the actual complexity of removing contents in decentralized platforms, the article concludes stressing the need of aligning the current regulatory framework to the metaverse and decentralized platforms, suggesting that new strategies and policies on content moderation within this peculiar digital environment should be discussed.

2. Content and conduct moderation in the EU framework

As it falls outside of the scope of the present contribution to focus on an in-depth review of the existing European Union legal framework on content moderation, it is sufficient to briefly recall the main regulation in the field, Regulation (EU) 2022/2065 (Digital Services Act – DSA), in order to better understand the current legal gaps when it comes to content moderation in the metaverse and in the decentralized internet.

In clearly stating that ‘Since August 2023, platforms have already started to change their systems and interfaces according to the Digital Services Act (DSA) to provide a safer online experience for all’⁷, the European Commission clearly sets and states the EU priority on content moderation policy and regulation, as to foster a safe (and democratic) internet for its users. The main goal of the DSA, that represents perhaps the most relevant legislation in the field of content moderation in the European Union, is to pose several obligations on internet intermediary services (art. 2.1) on removing illegal contents shared on the owned platform, clearly recalling its complementarity and non-prejudice to other EU acts in force in the field (explicitly mentioned under article 2.4, among which Directive 2010/13/EU; Union law on copyright and related rights; Regulation (EU) 2021/784 on the dissemination of terrorist content online; Regulation (EU) 2019/1148; Regulation (EU) 2019/1150; Regulations (EU) 2017/2394 and (EU) 2019/1020 and Directives 2001/95/EC and 2013/11/EU; Regulation (EU) 2016/679; and Directive 2002/58/EC).

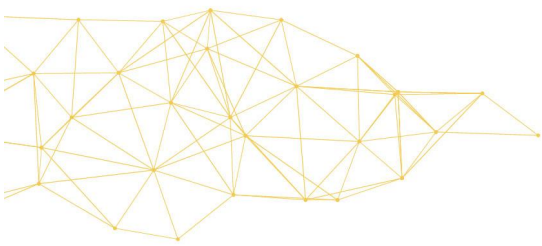
In consideration of the debate on the distinction between illegal and (otherwise) harmful content,⁸ it is worth stressing that the Digital Services Act poses clear obligations only with regard to illegal content, as defined both by Recital 12 (‘concept of “illegal content” should broadly reflect the existing rules in the offline environment’) and article 2, lit. h) (‘“illegal content” means any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law’), while service providers maintain an actual freedom to moderate harmful content, mostly depending on their policies and terms and conditions of use⁹.

The normative definition of content moderation is clearly stated under article 3.1(t) the DSA, as ‘the activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions [...]

⁷ See note 3

⁸ See note 5 for further clarification

⁹ An in-depth assessment on various platform’s terms and conditions is outlined in Britt van den Branden, Sophie Davidse, Eva Smit, ‘In between illegal and harmful: a look at the Community Guidelines and Terms of Use of online platforms in the light of the DSA proposal and the fundamental right to freedom of expression’ (*DSA Observatory*, 2 August 2021) <<https://dsa-observatory.eu/2021/08/02/in-between-illegal-and-harmful-a-look-at-the-community-guidelines-and-terms-of-use-of-online-platforms-in-the-light-of-the-dsa-proposal-and-the-fundamental-right-to-freedom-of-expression-part-1-of-3/>> accessed 28 June 2024



including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof [...]. Among the various obligations posed on service providers, it is useful to recall the prohibition on using deceptive and manipulative interfaces (called dark patterns, article 25); ensuring transparency behind recommender systems; restrictions on targeted advertising based on sensitive data (article 26) and information transparency behind the advertisements; removal of illegal contents (also following orders from national authorities); more generally, notification and action mechanisms, transparency about policies, terms and conditions, and finally obligations of reporting the steps and actions taken on content removal and moderation.

For the purposes of this paper, it is moreover sufficient to recall that, in short, while providers are generally not liable for contents published on a given platform by its users, the DSA requires providers to remove illegal content upon notice, being otherwise held liable if they fail in such removal having had actual knowledge of illegal activity or content, and if they do not act expeditiously to remove or disable access to the illegal content upon knowledge (article 6); on the other hand, general monitoring obligation on providers is excluded (so-called safe-harbour principle, under article 8), further liability exemptions applies in the cases outlined under article 4 in the cases of mere conduit¹⁰ and, finally, providers do not lose their liability privilege if they carry out voluntary investigations at their own initiative or take other measures aimed at detecting, identifying and removing or disabling illegal content (so-called Good Samaritan clause, under article 7).

When it comes to how a given content may be removed, the DSA does not require internet services to implement specific methods or tools of content screening and moderation – as long as moderation is efficient. Therefore, providers adopt several methodologies for content removal, such as automated means or manual interventions; differences in allowing or not allowing platform users' participation in the process via bottom-up reports or trusted flaggers (as Facebook or YouTube),¹¹ and even *ex ante* or *ex post* moderation policies. Nevertheless, given the scale and the speed at which such content is spread,¹² the vast majority of social media platforms implement AI-based solutions, in order to speed-up their detecting-and-removing pattern obligations at the largest possible extent.

Automated content moderation systems are filters based on algorithms and/or machine learning enabling a given system to employ automated filters aimed at removing contents that are pre-assessed as illegal by the algorithm (one above all: contents depicting child pornography).¹³ Although it has been stated that 'yet each platform keeps the specifics of how it enacts its moderation decisions opaque'¹⁴, what the DSA instead requires is that companies publish periodical transparency reports describing each platform's content moderation measures and information, including 'the error rate of

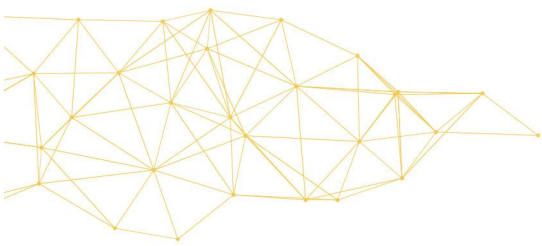
¹⁰ It refers to the liability privilege for providers that do not initiate the transmission; select the receiver of the transmission; or select and modify the information transmitted: this applies also in such cases when providers are aware of illegal content or activities.

¹¹ C.Drolsbach and N. Proelochs, 'Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database' (2023), <https://arxiv.org/html/2312.04431v1>; P. Friedl and J. Morgan, 'Decentralised content moderation' (2024) 13(2) Internet Policy Review

¹² M. Barral Martínez, 'Platform regulation, content moderation, and AI-based filtering tools: Some reflections from the European Union' (2023) 14 JIPITEC 211 para 1

¹³ T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (London, Yale University Press 2018)

¹⁴ S.Jhaver, I.Birman, E. Gilbert and A. Bruckman, 'Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator' (2019) 26(5) ACM Transactions on Computer-Human Interaction 1, 35



automated content moderation systems¹⁵ – with the rational of avoiding any form of machine learning bias that would make content moderation discriminatory and arbitrary. In short, the clear objective of the DSA is the prevention and suppression of illegal and harmful activities online, as well as the spread of disinformation, in order to ensure a user safe, fair, and open online environment, while preserving fundamental rights such as free speech and expression. Nevertheless, while the Regulation is surely to be considered a valuable legal tool, several questions on its future efficiency remain¹⁶, and concerns have been raised on its rigid structure¹⁷ and its vagueness when it comes to risk assessments.¹⁸ On top of this, its adequacy to target the intended objectives is yet to be assessed when it comes to facing the challenges arising from the increasing technological evolution.

3. The metaverse, the decentralized web and the issues for content moderation

The main challenges in the field of content moderation are hidden behind two emerging technologies capable of potentially changing the rules in the game of social networking and the internet: the metaverse and the decentralized web. While nothing suggests that, theoretically, the DSA would not apply, several issues arise. Let us first provide a technical overview and definition of these technologies, before analyzing the core of the problems.

Despite consensus still lacks around a universally adopted definition of the metaverse, also given the absence of legal provisions specifically concerning it, first it is essential to describe it as an online/digital platform with the features of realism, ubiquity, interoperability, and scalability¹⁹, as well as persistence and synchronicity²⁰. In addition, the metaverse is – or at least it is supposed to be – an immersive and constant environment, where virtual and tri-dimensional technological components, made accessible through virtual reality hardware,²¹ should make navigating in the metaverse an experience similar to real-life sensation. Nonetheless, the level of immersivity may vary depending on the use of virtual reality or, as it is more frequent at the current state of technological development, augmented reality components. In fact, while the latter implies a superposition of virtual objects on real-life scenarios or environments²², thus entailing a relative level of immersivity, a fully virtual environment would be capable of simulating real-

¹⁵ C. Drolsbach and N. Prollochs, *ibid.*

¹⁶ G. Miller, 'The Digital Services Act Is Fully In Effect, But Many Questions Remain' (*TechPolicy.press*, 20 February 2024) < <https://www.techpolicy.press/the-digital-services-act-in-full-effect-questions-remain/> > accessed 30 May 2024.

¹⁷ D. Keller, 'The DSA's Industrial Model for Content Moderation' (*Verfassungsblog: On Matters Constitutional*, 2022). doi:10.17176/20220224-121133-0

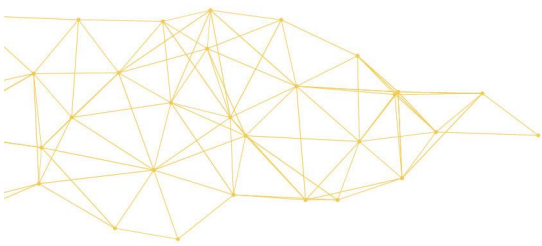
¹⁸ E. Douek, 'Content Moderation as Systems Thinking' (2022) 136(4) *Harvard Law Review*, doi:10.2139/ssrn.4005326

¹⁹ European Parliament, *Metaverse. Opportunities, risks and social implications* (2022). Specifically, realism means enabling 'people to become emotionally immersed in the virtual world'; ubiquity is the accessibility of the virtual spaces 'through all digital devices while using one virtual identity'; interoperability 'allows distinct systems or platforms to exchange information or interact with each other seamlessly' and finally scalability implies that the network architecture deliver sufficient power to enable massive numbers of users to occupy the metaverse without compromising the efficiency of the system and the experience of the users.

²⁰ M. Ball, 'The Metaverse: What It Is, Where to Find it, and Who Will Build It' (13 January 2020) <<https://www.matthewball.vc/all/themetaverse>>, accessed 5 May 2024

²¹ Council of the European Union, *Metaverse – Virtual World, Real challenges* (2022) Council of the European Union, Analysis and Research Team

²² R. T. Azuma, 'A Survey of Augmented Reality' (1997) 6(4) *Presence: Teleoperators and Virtual Environments*, MIT press 55, 385.



life sensations, especially when visual, tactile or even olfactory supporting devices are used.

Trying to combine all the abovementioned features, the metaverse can be defined as a tridimensional and relatively immersive digital platform, built with a non-necessarily contextual combination of augmented reality (AR), virtual reality (VR), mixed reality (MR), internet of things (IoT) and artificial intelligence (AI), enabling its users to have real life-experience²³.

For the sake of an exhaustive assessment, it is worth stressing that, as of today, most of the metaverse platforms appear somehow similar to more classical social media, in terms that most of the platforms are centralized and use augmented reality components – thus, the main difference relies on being the metaverse a relatively immersive and tri-dimensional environment. Therefore, the current use of the metaverse among the general public does not appear as enhanced yet and its development goes hand by hand with skepticism,²⁴ mostly because the current technological components of the metaverse, as well as the complex accessibility devices for the general public²⁵, does not enable fully immersive experiences yet.

Nevertheless, it is likely just a matter of time before technological progress will entail an enhanced use of virtual reality components (that along with time should also become more accessible in terms of costs): this will imply more intuitive and immersive experiences for metaverse users²⁶, where activities such as socializing, communicating, exchanging goods and services, making financial transactions, but also entertainment and other immersive experiences, will be extremely more realistic, so making the potential applications and uses of the metaverse likely to increase dramatically²⁷. Moreover, the actual impact of the metaverse on society and communication depends not only on how a given metaverse is built, whether with VR or AR components, but also on its architecture: in fact, while at the current technological development metaverses mostly present centralized infrastructures (this is the main feature of the so-called web 2.0, where a given platform is built and managed by a central node) the future of the internet and of metaverse

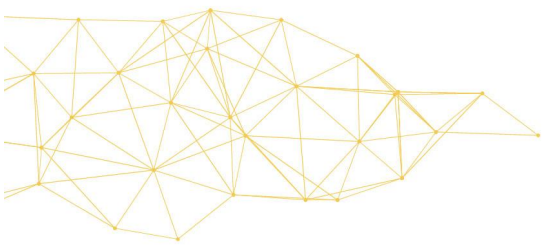
²³ A similar description can be read in Safari Kasiyanto, Mustafa R. Kilinc, 'The Legal Conundrums of the Metaverse' (2022) 1(2) Journal of Central Banking Law and Institutions 299, 322

²⁴ K. Wagner, 'Lessons From the Catastrophic Failure of the Metaverse' (*The Nation*, 3 July 2023) <<https://www.thenation.com/article/culture/metaverse-zuckerberg-pr-hype/>> accessed 29 May 2024.

²⁵ As of today, such hardware isn't that commonly used yet, nor are they comfortable to wear or economically sustainable David Chen, 'The Metaverse is Here... But is the Hardware Ready?' (*Spiceworks*, 14 March 2022, available at <<https://www.spiceworks.com/tech/hardware/guest-article/the-metaverse-is-here-but-is-the-hardware-ready/>> accessed 27 March 2023.

²⁶ European Commission, *Communication from the Commission to the European Parliament, the Council, The European Economic and Social Committee and the Committee of the Regions. An EU initiative on Web 4.0 and virtual worlds: a head start in the next technological transition* (2023) COM/2023 442/Final

²⁷ European Parliament, *Metaverse* (2023), Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies (PE 751.222), where it is clear stated that 'Yet, what is being developed and sold is a digital simulation, an information technology service depending on providers' IP address, servers and ICT services. In the future, general collective perception of metaverse as a reality may be reinforced if accorded with human natural perception in terms of high definition and multidimensionality of provided content and if its interface becomes more natural'.



platforms, accordingly, is possibly decentralized²⁸ – and this would represent ‘the soul of the metaverse’²⁹.

A decentralized architecture entails the absence of a central node (that is, for instance, internet intermediaries or any type of provider) controlling a network. Instead, data is processed and stored by multiple nodes as peer to peer networks and distributed ledgers such as the blockchain system.³⁰ The absence of central nodes owning a given network, that functions via the interaction of multiple private nodes working together, contextually de-powers internet service providers: this feature makes a decentralized web (or a decentralized platform) ‘... a system of interconnected, independent, privately owned computers that work together to provide private, secure, censorship-resistant access to information and services’,³¹ possibly resulting in a more democratic network that is less subject to external control or censorship, where users’ personal data are more resistant to privacy and security, and ensuring transparency and trustworthiness of online interactions (either simple communications or financial transactions) through the implementation of encryption mechanisms.

Summarizing, decentralized architectures, infrastructures or platforms imply an internet that is more private, secure, and transparent, where the absence of third parties holding powers make it independent and democratic, since ‘no central ownership, control, permission or possible censorship affects its users’³².

Currently, its main applications include data transfers, digital infrastructures and privacy, finance (transactions, smart contracts, cryptocurrencies, and NFTs, where the trust of the operation is given by the power of consensus mechanism)³³, machine learning such as natural language processing (NLP) or large language models (LLM), but it is likely that decentralized infrastructures will include a new typology of decentralized social media (such as the metaverses) and go as far as to fully blend physical and digital environments and landscapes, through the use of advanced intelligence, IoT, extended reality technologies and blockchain transactions.

²⁸ ‘Is decentralized internet the future?’ (*Security Senses*, 17 April 2024) <<https://securitysenses.com/posts/decentralized-internet-future>> accessed 30 April 2024; U.Dhariwal, ‘The Future of the Internet is Decentralized: Why Web3 Matters’ (*Medium*, 21 March 2024) <<https://ujjwaldhariwal0.medium.com/the-future-of-the-internet-is-decentralized-why-web3-matters-9ab5ac8f5056>> accessed 25 April 2024.

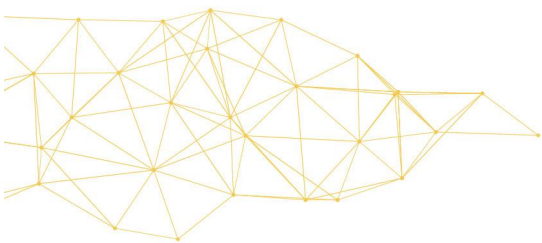
²⁹ T. Huynh-The et al., ‘Blockchain for the metaverse: A review’ (2023) 143 *Future Generation Computer Systems* 401, 419

³⁰ A. Bhalla, ‘Decentralized vs. Centralized Digital Networks: Understanding the differences’ (*Blockchain Council*, 10 May 2024) <<https://www.blockchain-council.org/blockchain/centralized-vs-decentralized-digital-networks/#:~:text=Centralized%20networks%20are%20owned%20and,of%20by%20a%20single%20author%20ity>>, accessed 30 May 2024.

³¹ F. Aboukhadijeh, ‘What is the Decentralized Web? 25 experts break it down’ (Syracuse University) <<https://onlinegrad.syracuse.edu/blog/what-is-the-decentralized-web/#:~:text=%E2%80%9CThe%20term%20'Decentralized%20Web',%20factor%20control%20or%20censorship.%E2%80%9D>> accessed 5 May 2024.

³² S. Shilina, ‘The future of social networking: Decentralization for user empowerment, privacy, and freedom from censorship’ (*Medium*, 6 November 2023) < <https://medium.com/paradigm-research/the-future-of-social-networking-decentralization-for-user-empowerment-privacy-and-freedom-from-a0a8f74790cb>> accessed 3 May 2024.

³³ Ch. Chen, Lei Zhang, Yihao Li, Tianchi Liao, Siran Zhao, Zibin Zheng, Huawei Huang, Jiajing Wu, ‘When Digital Economy Meets Web 3.0: Applications and Challenges’ (2022) PP(99) *IEEE Open Journal of the Computer Society* 1, 12. In their article, the authors provide a comparison between web 3.0 and web 1.0 and web 2.0, in terms that the web 1.0 has the features one way information, professional generated content, read-only and portal internet, centralization, while web 2.0 (current version of the web) has interactive information, user generated content, read-and-write interactive internet, centralization.



Once highlighting the main technical features of the metaverse and the decentralized web, let us assess the issues arising in the field content moderation accordingly. Concerning the metaverse, the issues focus either on the typology of contents that users might generate in the metaverse, potentially implying a re-assessment of which contents should be moderated or removed, and on the monitoring and detection of illegal contents. Concerning the decentralized web, the main problem consists in the practical complexity of moderating contents in social networks built with decentralized architectures, given the de-powered role of central nodes in controlling the network.

3.1. Revisiting content moderation in light of virtual environments

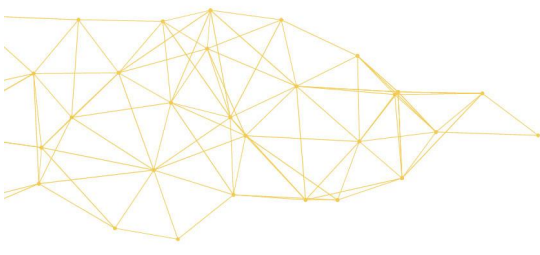
Content moderation in the metaverse raises two orders of challenges: a) the first, more substantial, concerns the typologies of what can be considered as an illegal or harmful content³⁴ to be moderated and eventually removed. Contrarily to bi-dimensional contents such as comments, posts, pictures, or other USGs, the tri-dimensional feature of the metaverse enriches the range of potentially illegal conducts: as immersive environments are capable of giving a real-life feel to their users, typologies of unwanted behaviours might dramatically increase, also in different forms; b) the second issue is more procedural (or practical, so to say) and it concerns the detection, screening and moderation of illegal conducts: since the metaverse is a synchronous environment, most of the content moderation tools and policies implemented so far (including automated systems, see below paragraph 3.2) may not be effective, as synchronicity represents a game changer when it comes both to prevention and to repression of illegal contents.

Concerning the first issue, it is useful to recall a recent document issued by the Interpol that taxonomizes the potential harms in the metaverse.³⁵ For the purposes of this paper, crimes committed in the metaverse can be divided in two typologies: the first typology includes the crimes that can be considered as illegal content both in bi-dimensional platforms and in the metaverse: cybercrimes (unlawful access, doxing, ransomware, data theft), crimes related to sexual offenses or abuse (stalking, child pornography, harassment, exploitation or indecent exposure), acts causing emotional distress (defamation, libel, non-consensual sharing of intimate images), crimes against public safety (mis- and dis-information, propaganda, drug trafficking), identity crimes (hacking, violation of privacy, identity theft), terrorism-related crimes (financing, radicalization) or financial crimes (fraud, scams, money laundering, phishing). Here, the bi- or tri-dimensional feature of the environment does not change the potential harm caused by a conduct on the victim: in both cases, the conduct is to be considered unlawful (and the content moderated and removed).

The second group of crimes includes all those contents gaining an enhanced level of harmfulness in reason of the virtual element underlying them. Such conducts would not be considered illegal (or even harmful) in bi-dimensional platforms, since augmented/virtual reality is essential to make the conduct actually harmful: without the virtual element, there wouldn't be a crime and, as such, these conducts can be committed only in metaverses. Let us think about several forms of property crimes (trespassing in private virtual space, virtual burglary, theft of virtual property or assets, robbery from an

³⁴ Bearing in mind the debate outlined above (see note 5).

³⁵ Interpol, *Metaverse: A law enforcement perspective. Use cases, crime, forensics, investigations, and governance* (White Paper, January 2024)



avatar), physical crimes (sexual virtual abuse and assault, virtual rape – of which multiple reports have already been filed)³⁶, or peculiar intellectual property or copyright infringement crimes (in which the virtual element enhances the features of a given counterfeit), but also of actual virtual-physical attacks occurring on avatars. This group does not correspond to any other conduct in traditional and bi-dimensional social media.

The virtual component of the metaverse enables a real-life experience that is not inherently proper of bi-dimensional platforms and, as such, is capable of causing major emotional distress and harm: recent researches highlight that the effect of considering an avatar as the virtual representation/extension of its own self is common among metaverse users³⁷ and, as such, ‘Even though these experiences are virtual, they can have tangible psychological effects on victims, including trauma, fear, and distress, mirroring the impacts of real-world sexual violence’.³⁸ Although a general consensus lacks when it comes to considering a virtual contact as an actual physical contact, what it is sure is that there is evidence that virtual contact is capable of reflecting on psychological harm, sometimes to an extent as to not distinguishing between virtual and real life.³⁹

Summarizing, the metaverse entails new typologies of harmful conducts which, just like the traditional and bi-dimensional illegal contents to be moderated, should be removed in order to ensure a safe environment for metaverse users.

The second issue related to content moderation in the metaverse embodies a practical nature of crime prevention and law enforcement: the challenge is how to detect, screen and eventually remove conducts or behaviours that are synchronous, and the issue is even exacerbated when we consider the scale and the volume of content in a metaverse – that could potentially ‘host countless virtual environments and interactions’ and thus ‘exponentially increases the volume of content requiring moderation’.⁴⁰

Solutions so far have focused on synchronous moderators ensuring safe environments basically constantly navigating in metaverses and trying to prevent and detect unlawful behaviours of avatars – given the fact that ‘traditional moderation tools, such as AI-enabled filters on certain words, don’t translate well to real-time immersive

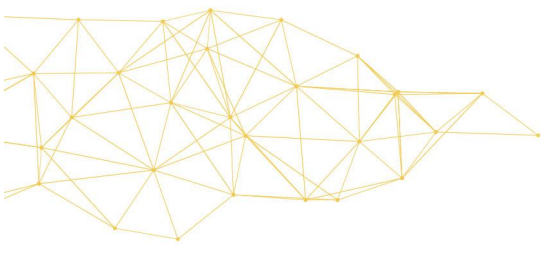
³⁶ Several articles can be found on people reporting having suffered sexual violence in the metaverse. See <https://www.repubblica.it/esteri/2024/01/07/news/gran_bretagna_stupro_metaverso_indagini-421824572/>,<<https://www.dailymail.co.uk/news/article-12917329/Police-launch-investigation-kind-virtual-rape-metaverse.html>>,<<https://tg24.sky.it/mondo/2024/01/06/abusi-sessuali-metaverso-regno-unito>>,<https://www.eko.org/images/Metaverse_report_May_2022.pdf>,<<https://www.independent.co.uk/tech/rape-metaverse-woman-oculus-facebook-b2090491.html>>,<<https://www.cnbctv18.com/technology/woman-recalls-gang-rape-in-metaverse-concerns-grow-over-making-vr-platforms-safe-from-sexual-predators-12396992.htm>>

³⁷ Guo Freeman, Samaneh Zamanifard, Divine Maloney and Alexandra Adkins, ‘My Body, My Avatar: How People Perceive Their Avatars in Social Virtual Reality’ (2020) Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems 1,8 <<https://doi.org/10.1145/3334480.3382923>>

³⁸ Mohamed Chawki, Basu Subhajit and Choi Kyung-Shick, ‘Redefining Boundaries in the Metaverse: Navigating the Challenges of Virtual Harm and User Safety’ (2024) 13(3) *Laws* 33, <<https://doi.org/10.3390/laws13030033>> accessed 27 May 2024

³⁹ Thomas D. Parsons, Christopher Courtney, Louise Cosand, Arvind Iyer, Albert A. Rizzo and Kelvin Oie, ‘Assessment of Psychophysiological Differences of West Point Cadets and Civilian Controls Immersed within a Virtual Environment’ in Dylan D. Schmorow, Ivy V. Estabrooke and Marc Grootjen (eds.), *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience* (Lecture Notes in Computer Science. Berlin, Heidelberg, 514–23, 2009)

⁴⁰ Michelle Teo, ‘The Wellbeing of Content Moderators in the Metaverse: Navigating New Realities’ (*ZevoHealth*, 21 March 2024) <<https://www.zevohealth.com/blog/the-wellbeing-of-content-moderators-in-the-metaverse-navigating-new-realities/>>



environments'.⁴¹ But also whether we consider automated systems efficient to detect virtual interaction, this would imply that 'every second of every interaction to be monitored and analyzed [...] it would require untold amounts of computing power, making it practically impossible'.⁴²

Other researches highlighted that content moderation in the metaverse is not necessarily synchronous, thus moderators' real-time immersion is not necessary: virtuous examples as Roblox moderation strategies, for instance, show how avatars' interactions can be evaluated *ex post* via innovative technological tools enabling user-to-user moderation in near real-time, underscoring 'a shift towards more nuanced and technologically aided moderation strategies, effectively balancing human oversight with automated processes to maintain a safe and inclusive digital environment'.⁴³ According to this literature, the issue of moderation would be practically solved with AI-based tools, where the actual challenges concerning content moderation in the metaverse would rather focus on avatars' anonymity and identities, the scale and volume of the contents, the evolving typologies of contents (as also highlighted above) as well as legal and ethical uncertainties related with the lack of clear regulatory definitions of the metaverse and its users' activities.⁴⁴

As showed, uncertainty underlies content moderation in the metaverse, both under a substantial standpoint – where the typologies of illegal and harmful contents/conducts can be exponentially larger due to the immersive and virtual nature of the environment, thus including those behaviours that would be harmless (and would not even exist in practice) in bi-dimensional social media; and under a practical standpoint, where the vast scale and the synchronous feature of the metaverse would make it extremely complex to detect (near) real-time behaviours if moderators do not have an advanced technological equipment detecting illegal conducts – and even then, other issues still underlie content moderation.

3.2. Content moderation in decentralized platforms

Internet intermediaries or service providers have to fulfill obligations associated with content moderation either in centralized and decentralized internet or platforms – as current regulations do not provide any distinction. Nonetheless, when it comes to decentralized architectures the fulfillment of such obligations may result (more than) complex: given a network's decentralized structure, where central nodes are basically depowered, central governance mechanisms lack control⁴⁵ and the practical removal of a given content is considered to be potentially even impossible.⁴⁶

⁴¹ Tate Ryan Mosley, 'How an undercover content moderator polices the Metaverse' (*MIT Technology Review*, 28 April 2023) <<https://www.technologyreview.com/2023/04/28/1072393/undercover-content-moderator-polices-the-metaverse/>>

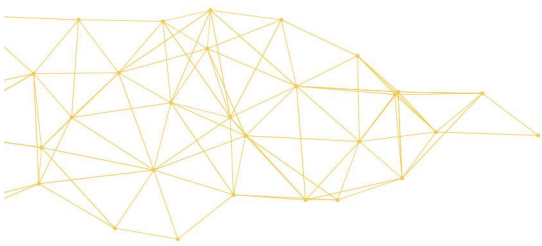
⁴² Ryan Hsu, 'Meet the new 'verse, same as the old 'verse: Moderating the Metaverse' (*Georgetown Law Technology Review*, May 2022) <<https://georgetownlawtechreview.org/meet-the-new-verse-same-as-the-old-verse-moderating-the-metaverse/GLTR-05-2022/>>

⁴³ Michelle Teo, *ibid.*

⁴⁴ Michelle Teo, *id.*

⁴⁵ Adi Robertson, 'How the Biggest Decentralized Social Network Is Dealing with Its Nazi Problem' (*VERGE*, 12 July 2019) <<https://perma.cc/QA6F-J54U>> accessed 24 May 2024.

⁴⁶ Adam Hadley, 'Terrorists Will Move to Where They Can't Be Moderated' (*Wired UK*, 31 May 2021) <<https://www.wired.co.uk/article/terrorists-dweb>> accessed 14 May 2024; Bennett Clifford, 'Moderating



As stated above, the existing literature in the field highlights the beneficial effects of decentralized networks because they imply a censorship-resistant system, given the 'distributed content storage, community-driven moderation, and anti-censorship tools' where it 'becomes immutable and impervious to alteration or removal by any central authority'⁴⁷, thus fostering free expression and free speech.⁴⁸ Several decentralized and blockchain social networks are proliferating, enhancing security and users' control over central entities and big techs control and power. In a notable contribution, Shilina⁴⁹ enlightens the state-of-art of such decentralized platforms including Peepeth, Minds, Mirror, Sapien, CyberConnect or Status; but also traditionally-built Web2 platforms such as Reddit or Twitter (X), that introduced several decentralized supports for their users (subreddit communities for Reddit, or NFT for Twitter). In short, it is undoubtable that decentralized social media ensure less central control by de-powering central entities, appear more user-friendly, democratic and somehow more fair, making decentralized content moderation a 'potential response to many of the issues linked to the current prevalent models in the advertisement-driven attention economy'⁵⁰ by ensuring a more participatory digital governance⁵¹ and, of course, de-powering uncontrolled moderation⁵², especially when it comes to arbitrary criteria of moderation transcending the harmful content.

Nevertheless, not only content moderation obligations still exist – and do represent a challenge – but leaving moderation basically unregulated could be dangerous for the online community, in terms of presence of illegal or harmful content. Several mechanisms have been implemented so far and could be used to solve this conundrum, but all of them bring along problems, as also highlighted by the existing literature.

The first includes the so-called federated networks, that are defined as peer-to-peer and open source networks operating via multiple coordinated and interoperable decentralized nodes. Despite the networks being managed by a central management entity, this does not represent its core and does not operate as a central internet provider.⁵³ Here, despite the presence of a single entity, 'management, control, and data plans are distributed over multiple networks or locations'⁵⁴. Its application can include social media federated networks such as Mastodon⁵⁵, that operates with the ActivityPub protocol: here, each developer is free to run its independent and autonomous server, so-called instance, where each instance is interconnected and has to follow the central

Extremism: The State of Online Terrorist Content Removal Policy in the United States' (*George Washington Program on Extremism*, December 2021) <<https://perma.cc/3JG8-KSTV>> accessed 24 May 2024.

⁴⁷ Sasha Shilina, *ibidem*

⁴⁸ Mike Masnick, 'Protocols, not platforms: A technological approach to free speech' (*Knight First Amendment Institute*, 21 August 2019) <<https://perma.cc/MBR2-BDNE>> 20 May 2024.

⁴⁹ Sasha Shilina, *ibidem*

⁵⁰ Paul Friedl and Julian Morgan, *ibid.*

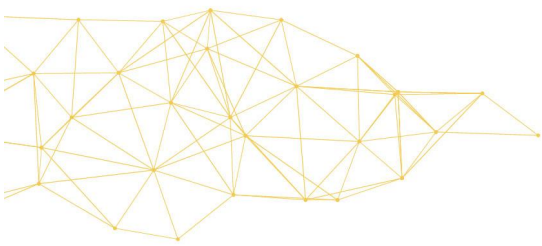
⁵¹ Eleonora Bietti, 'A genealogy of digital platform regulation' (2023) 7 *Geo L. Tech. Rev*

⁵² Ksenia Ermoshina and Francesca Musiani, 'Safer spaces by design? Federated architectures and alternative socio-technical models for content moderation' (2022), Annual Symposium of the Global Internet Governance Academic Network (GigaNet) <<https://hal.science/hal-03930548/document>>

⁵³ 'Federated Architecture: learn the Key benefits and how it is different from other architectures' (*Atlan*, 31 August 2023) <<https://atlan.com/federated-architecture/#what-is-a-federated-architecture>>

⁵⁴ 'What is a federated network' (*VMWare*) <<https://www.vmware.com/topics/glossary/content/federated-network.html>>

⁵⁵ Matteo Zignani, Christian Quadri, Sabrina Gaito, Hocine Cherifi, Gian Paolo Rossi, 'The Footprints of a "Mastodon": How a Decentralized Architecture Influences Online Social Relationships' (2019) IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) 472, 477



protocol's terms and conditions – including content policies.⁵⁶ While on one hand this features embody the positive outcome of creating new interaction patterns,⁵⁷ on the other hand the lack of central control implies that 'there is no central authority that can decide which instances are valid or that can ban a user or a piece of content from the ActivityPub network entirely'.⁵⁸ Surely, it is true that decentralized platforms provide terms and conditions to navigate and share contents – it is the case of Mastodon where, despite the numerous and autonomously regulated instances, its terms and conditions encourage shared content moderation policies among its instances⁵⁹ – and each instance requires to follow such terms and conditions but, nevertheless, this represents a form of non-binding regulation and, absent the possibility of controlling or removing decentralized instances in terms of law enforcement, it results basically inefficient.

Finally, while it is true that in federated systems an illegal instance (or user) can be shut down and blocked by other instances (or users), this solution would work only with a general consensus of all the other instances – and the question would be on which grounds to determine an aprioristic general consensus of all the instances.

The second mechanism is community-driven decentralized content moderation, which includes those social media that enable their users to report, flag or, more generally, participate in the moderation process. This is the case of Reddit, which, despite being a web.2-based company, enables its subreddits to have a certain level of autonomy in moderating content on a bottom-up approach. Nevertheless, here the risk is that community-driven decentralized content moderation actually masks a centralized content moderation: taking the example of Reddit, it can eventually modify/take the final decision after Subreddit operations. This implies that there still exist central entities eventually moderating a given content if dissatisfied with community-driven decentralized operations: as such, the democratic aspect of decentralized platforms would no longer exist, content moderation mechanisms would not present any difference when compared to centralized platforms and the system would no longer be censorship-resistant.

A third option could be the implementation of technology-driven systems enabling automated control, which would screen contents and detect (and also prevent) the share of illegal content. This is the case of Microsoft PhotoDNA (whose mechanism is mostly used to detect child pornography)⁶⁰ and other AI-driven systems used both in centralized and decentralized platforms. Nevertheless, such automated control would work only if a decentralized platform is enough speedy, scalable, and has enough economic resources to bear the costs of automated systems, otherwise being inefficient. Moreover, at the current

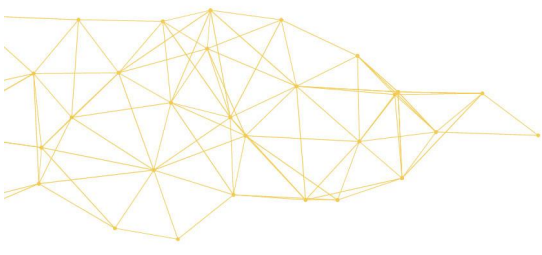
⁵⁶ Christopher Lemmer Webber, 'Mastodon launches their ActivityPub support, and a new CR!' (ActivityPub, 10 September 2017), < Mastodon launches their ActivityPub support, and a new CR! -- ActivityPub Rocks!>

⁵⁷ Diana Zulli, Miao Liu and Robert Gell, 'Rethinking the "social" in "social media": insights into topology, abstraction, and scale on the Mastodon social network' (2020) 22(7) *New Media & Society* 1188, 1205

⁵⁸ Alan Z. Rozenshtein, 'Moderating the Fediverse: Content Moderation on Distributed Social Media' (2022) 3 *Journal of Free Speech Law* 217, Minnesota Legal Studies Research Paper No. 23-19 <SSRN: <https://ssrn.com/abstract=4213674>> accessed 30 May 2024

⁵⁹ As stated in Rozenshtein, *id*: 'Specifically, the Mastodon project has promulgated a "Mastodon Server Covenant," whereby instances that commit to "[a]ctive moderation against racism, sexism, homophobia and transphobia" such that users will have "confidence that they are joining a safe space, free from white supremacy, anti-semitism and transphobia of other platforms" are eligible to be listed on the project's homepage as recommended instances. See Eugen Rochko, *Introducing the Mastodon Server Covenant, MASTODON* (May 16, 2019), <https://perma.cc/GP8H-MXXX>. But the covenant is not binding on any Mastodon instance, and non-complying instances remain full-fledged member of the overall Mastodon network, subject only to the moderation decision of other instances'.

⁶⁰ Hany Farid, 'Reining in Online Abuses' (2018) 19 *Tech & Innovation* 596



state of technological development not all automated systems are capable of detecting all the different typologies and forms of illegal user generated content.

Fourth, local legislations criminalizing the spread of illegal content on the internet would ensure law enforcement, turning down a given instance as a whole and/or prosecuting its members. While this would surely be a secure and efficient tool, relying on the mere law enforcement and subsequent prosecution of illegal instances' owners' does not necessarily prevent the spread of such contents. Moreover, certain conducts might be considered illegal in one country but not in another and, in terms of conflict of law and jurisdiction, this might raise serious issues.⁶¹

Finally, every user of a decentralized platform should consent and agree to comply with a given platforms' terms and conditions and/or terms of use. Here, as already stressed, in case of non-compliant users, internet providers in decentralized networks lack the practical power to moderate and eventually remove a given content. Therefore, law enforcement agencies should remove illegal contents, but this does not look like an implementable option as of today: in fact, agencies' need to speed up in understanding new forms of crimes emerging from nascent technologies [...] law enforcement may not be able to deal with [...] which pose great challenges to law enforcement due to the potential ability to run without an accountable authority'.⁶²

All the abovementioned issues might not appear as an urgent challenge today, when the use of decentralized networks and systems can still be considered marginal among the general public. But what if – and apparently, it's a matter of *when*, not of *if* – the use of the Web3 and decentralized social media will become an everyday activity for millions, if not billions of people? How would scalability-related issues be solved? Will automated systems be trained enough to detect all the potential illegal activities? Finally, will law enforcement have enough capacity to replace providers' content moderation duties, given the practical impossibility to moderate decentralized networks in absence of a community-driven approach? Will tailored legal solutions be implemented?

Absent clear legal provisions and largely efficient best practices, as of today the questions on a widespread adopted solution for decentralized content moderation remain open.

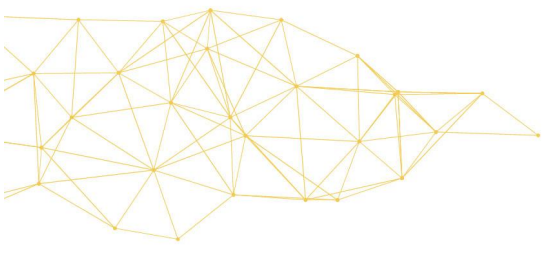
4. Content moderation in decentralized metaverses

Before concluding, let us briefly put together content moderation issues in the metaverse and content moderation issues in decentralized platforms. While as we have seen, both autonomously imply critical challenges to traditional content moderation (obligations), to the extent of the possible inefficiency of the current regulations in the field. Such challenges are even exacerbated when it comes to decentralized metaverses (such as Decentraland, Sandbox or Roblox), where the issues related to the extent of what can be considered as an illegal or harmful content in the metaverse, come together with the issues related to content moderation in decentralized platforms.

The metaverse market is expected to reach USD 426.9 billion by 2027, dramatically increasing flows of capitals compared to the USD 61.8 billion in 2022, and expanding at a

⁶¹ For the purposes of this paper, the issue of jurisdiction in online environments is not covered, as it would imply an in-depth analysis falling outside its scope.

⁶² Interpol, *ibid.*



CAGR of 47.2%⁶³. An essential role is played by the blockchain technology underlying decentralized ecosystems, as it would foster the increased use of artificial intelligence and haptic technologies in virtual environments in order to improve users' immersive experiences and generate engaging content.⁶⁴ Accordingly, the more engaging and immersive the experience is, the higher the number of users expected to join it.

In possibly being the future of the social networks, decentralized metaverses will surely represent the most democratic, accessible and community driven form of digital interaction, lacking the control and the potential risks of censorship of central entities. On the other hand, as highlighted, decentralized metaverses will present new typologies and forms of user generated content, ranging from traditional speech or visual elements, to actual virtual/physical conducts: this represents a challenge for content moderation, since what will be considered as an illegal or harmful content is yet to be precisely assessed. Moreover, screening and control operations are made difficult by the synchronous feature of the metaverse, in which avatars act in real-time. On top of this, we have seen the complex challenges surrounding both content moderation in decentralized platforms and the critical observations that can be raised on decentralized forms of content moderation – such as federated networks, automated systems and community based or law enforcement solutions.

Summing it up: given the novelty of the issue and absent being regulations specifically focusing on content moderation, a malicious exploitation of decentralized metaverses could pose threats not only on individual safety (as widely seen above), but also on national security: although the use of decentralized platforms has been proved to be still marginal in this field,⁶⁵ the metaverse and the virtual worlds represent a growing concern⁶⁶ as their widespread use may foster terrorist propaganda, recruitment, combat training and coordination and organization of attacks.⁶⁷ On top of this, illegal activities such as money laundering, illegal financing and fundraising would be difficult to prevent, detect and prosecute when committed in decentralized environments.⁶⁸

In short, decentralized metaverses represent a future issue – perhaps, the most challenging one when it comes to content moderation in digital environments.

5. Conclusions

The present paper provided a brief overview of content moderation obligations for internet intermediary services, highlighting the current regulations in force at the European Union level. A deeper study of the issue suggests that, while nothing suggests that the current legal framework could not theoretically apply to metaverse providers and decentralized platforms, several issues arising from the mentioned new technologies might make it less (if not at all) efficient.

⁶³'10 Best Metaverse platforms to watch out in 2023' (*Blockchain Council*, 27 September 2023) <<https://www.blockchain-council.org/metaverse/10-best-metaverse-platforms/>>

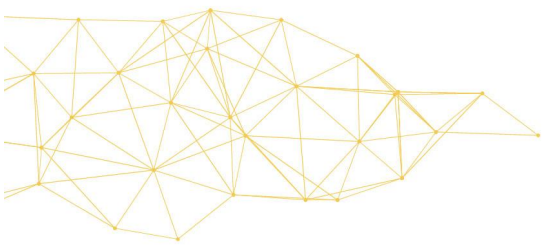
⁶⁴ *Id.*

⁶⁵Lorand Bodo and Inga K. Trauthig, 'Emergent Technologies and Extremists: The DWeb as a New Internet Reality?', (*Global Network on Extremism and Technology*, ICSR, King's College, London, 2022) <<https://gnet-research.org/wp-content/uploads/2022/07/GNET-Report-Emergent-Technologies-Extremists-Web.pdf>>

⁶⁶Mauro Miedico, 'The Application of Augmented Reality and Virtual Reality Technologies in Countering Terrorism and Preventing Violent Extremism', (*United Nations Office for Counter Terrorism*, 8 July 2021)

⁶⁷EU Counter-Terrorism Coordinator, 'The Metaverse in the context of the fight against terrorism' (Report n. 9292/2022).

⁶⁸ Interpol, *ibid.*



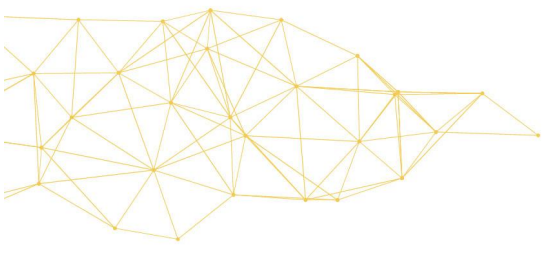
Specifically, on one hand content moderation in the metaverse appears peculiar both under a substantial (in terms of typologies of illegal or harmful content, given its immersive nature) and procedural (in terms of screening, evaluation and moderation of such content, given its synchronous nature) standpoints; on the other hand, content moderation in decentralized platforms is by its own nature non-implementable by central nodes and thus, in absence of community-based shared – and followed – policies, illegal contents might not be removed at all. Moreover, when the abovementioned peculiarities get together in decentralized metaverses, the risk of practical impossibility of moderating and removing harmful contents online represents more than a threat and, given the expected metaverse market within a few years, the need of a solution appears urgent.

While the Digital Services Act represents an efficient tool for the current internet and social networks, future challenges might urge its revision in the terms better explained above. At the same time, scientific literature on the issue appears still scarce. Therefore, the general need of discussing new strategies and policies to address the issues, as well as the urge of shaping legal frameworks capable of regulating it, is pivotal.

In this regard, a possible solution – aimed at least at preventing the spread of illegal contents – could be the implementation of mechanisms that, instead of trying to punish and/or sanction those who misbehave (given its practical complexity in decentralized architectures), are aimed at rewarding those who don't misbehave.⁶⁹ For instance, virtuous behaviours such as generating legal or at least neutral contents, or reporting potentially harmful contents, could be evaluated by other platform users via trusted positive flags (let us call it a 'green flag' for virtuous behaviours). This mechanism would represent compensation for virtuous users, with incentives in a given platform (e.g., tokens for purchases, access to additional benefits, and similar mechanisms) and, therefore, misbehavior would be discouraged. Of course, this solution could work only whether the platform providers cooperate in providing such rewards.

In conclusion, a further question remains open for – and to be considered by – the policymakers: in light of the evolving technologies that appear to outdate freshly issued legal frameworks in the spectrum of a few years, can we re-think the ways of making legislation, namely, can we move from the traditional form? A question that involves ethics and philosophy of law, that cannot be answered today but urges further research and discussion.

⁶⁹ A mechanism that would not be dissimilar from the Mining Reward in Bitcoin and cryptocurrencies systems (<https://crypto.com/glossary/it/mining-reward>)



Bibliography

Aboukhadijeh F, 'What is the Decentralized Web? 25 experts break it down' (Syracuse University) <<https://onlinegrad.syracuse.edu/blog/what-is-the-decentralized-web/#:~:text=%E2%80%9CThe%20term%20'Decentralized%20Web',%2Dactor%20control%20or%20censorship.%E2%80%9D>> accessed 5 May 2024

Azuma R T, 'A Survey of Augmented Reality' (1997) 6(4) Presence: Teleoperators and Virtual Environments, MIT press 55, 385

Ball M, 'The Metaverse: What It Is, Where to Find it, and Who Will Build It' (13 January 2020) <<https://www.matthewball.vc/all/themetaverse>>, accessed 5 May 2024

Barral Martínez M, 'Platform regulation, content moderation, and AI-based filtering tools: Some reflections from the European Union' (2023) 14 JIPITEC 211 para 1

Bhalla A, 'Decentralized vs. Centralized Digital Networks: Understanding the differences' (Blockchain Council, 10 May 2024) <<https://www.blockchain-council.org/blockchain/centralized-vs-decentralized-digital-networks/#:~:text=Centralized%20networks%20are%20owned%20and,of%20by%20a%20single%20authority>>, accessed 30 May 2024

Bietti E, 'A genealogy of digital platform regulation' (2023) 7 Geo L. Tech. Rev

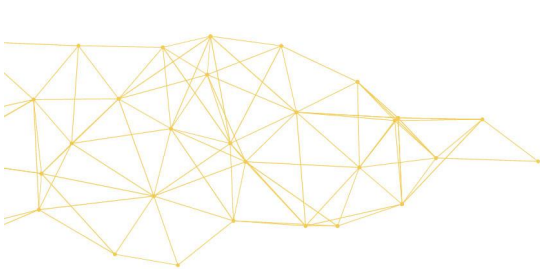
Bodo L and Trauthig I K, 'Emergent Technologies and Extremists: The DWeb as a New Internet Reality?', (Global Network on Extremism and Technology, ICSR, King's College, London, 2022) <<https://gnet-research.org/wp-content/uploads/2022/07/GNET-Report-Emergent-Technologies-Extremists-Web.pdf>>

Chen C, Zhang L, Li Y, Liao T, Zhao S, Zheng Z, Huang H, Wu J, 'When Digital Economy Meets Web 3.0: Applications and Challenges' (2022) PP(99) IEEE Open Journal of the Computer Society 1, 12

Chen D, 'The Metaverse is Here... But is the Hardware Ready?' (Spiceworks, 14 March 2022, available at <<https://www.spiceworks.com/tech/hardware/guest-article/the-metaverse-is-here-but-is-the-hardware-ready/>> accessed 27 March 2024

Chawki M, Subhajt B and Kyung-Shick C, 'Redefining Boundaries in the Metaverse: Navigating the Challenges of Virtual Harm and User Safety' (2024) 13(3) Laws 33, <<https://doi.org/10.3390/laws13030033>> accessed 27 May 2024

Clifford B, 'Moderating Extremism: The State of Online Terrorist Content Removal Policy in the United States' (2021), Program on Extremism. George Washington University, <<https://extremism.gwu.edu/sites/g/files/zaxdzs5746/files/Moderating%20Extremism%20The%20State%20of%20Online%20Terrorist%20Content%20Removal%20Policy%20in%20the%20United%20States.pdf>> accessed 27 April 2024



Commission of the European Communities, Communication from the Commission to the Council, the European Parliament, the Council, The Economic and Social Committee and the Committee of the Regions. Illegal and harmful content on the internet (1996) COM(96) 487/Final

Council of the European Union, Analysis and Research Team Metaverse – Virtual World, Real challenges (Report, 2022)

Dhariwal U, 'The Future of the Internet is Decentralized: Why Web3 Matters' (Medium, 21 March 2024) <<https://ujjwaldhariwal0.medium.com/the-future-of-the-internet-is-decentralized-why-web3-matters-9ab5ac8f5056>> accessed 25 April 2024.

Drolsbach C and Prollochs N, 'Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database' (2023), <https://arxiv.org/html/2312.04431v1>

Douek E, 'Content Moderation as Systems Thinking' (2022) 136(4) Harvard Law Review, doi:10.2139/ssrn.4005326

Ermoshina K and Musiani F, 'Safer spaces by design? Federated architectures and alternative socio-technical models for content moderation' (2022), Annual Symposium of the Global Internet Governance Academic Network (GigaNet) <<https://hal.science/hal-03930548/document>>

European Commission, Communication from the Commission to the European Parliament, the Council, The European Economic and Social Committee and the Committee of the Regions. An EU initiative on Web 4.0 and virtual worlds: a head start in the next technological transition (2023) COM/2023 442/Final

European Commission, 'The impact of the Digital Services Act on digital platforms' (European Union, 30 April 2024) <<https://digital-strategy.ec.europa.eu/en/policies/dsa-impact-platforms>> accessed 21 May 2024

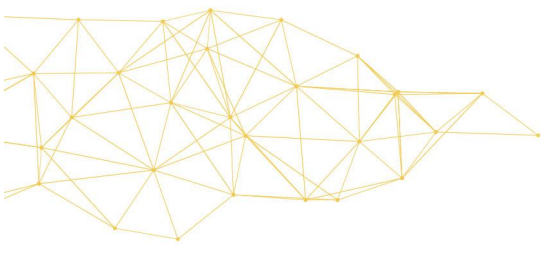
European Parliament, Metaverse (2023), Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies (PE 751.222)

European Parliament, Reform of the EU liability regime for online intermediaries. Background on the forthcoming digital services act (2020), European Parliamentary Research Service (PE 649.404)

European Parliament, Metaverse. Opportunities, risks and social implications (Report, 2022).

EU Counter-Terrorism Coordinator, 'The Metaverse in the context of the fight against terrorism' (Report n. 9292/2022)

EU Counter-Terrorism Coordinator, 'Online gaming in the context of the fight against terrorism' (Report n. 9066/2020)



Farid H, 'Reining in Online Abuses' (2018) 19 Tech & Innovation 596

Flew T, Martin F and Suzor N, 'Internet regulation as media policy: rethinking the question of digital communication platform governance' (2019), 10(1) Journal of Digital Media & Policy 33, 50.

Freeman G, Zamanifard S, Maloney D and Adkins A, 'My Body, My Avatar: How People Perceive Their Avatars in Social Virtual Reality' (2020) Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems 1,8 <<https://doi.org/10.1145/3334480.3382923>>

Friedl P and Morgan J, 'Decentralised content moderation' (2024) 13(2) Internet Policy Review

Gillespie T, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (London, Yale University Press 2018)

Grant J I, 'Removing the risks from a decentralised internet' (The Strategic, 2021) Australian Strategic Policy Institute, 30 July 2021, <https://www.aspistrategist.org.au/removing-the-risks-from-a-decentralised-internet/>

Hadley A, 'Terrorists Will Move to Where They Can't Be Moderated' (Wired UK, 31 May 2021)

<<https://www.wired.co.uk/article/terrorists-dweb>> accessed 14 May 2024

Hsu R, 'Meet the new 'verse, same as the old 'verse: Moderating the Metaverse' (Georgetown Law Technology Review, May 2022) <<https://georgetownlawtechreview.org/meet-the-new-verse-same-as-the-old-verse-moderating-the-metaverse/GLTR-05-2022/>>

Huynh-The T et al., 'Blockchain for the metaverse: A review' (2023) 143 Future Generation Computer Systems 401, 419

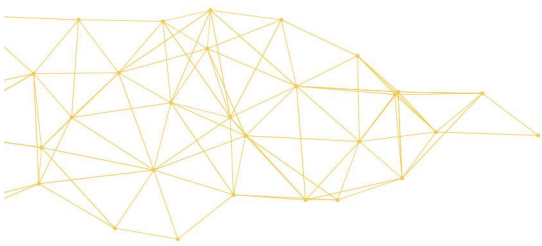
Interpol, Metaverse: A law enforcement perspective. Use cases, crime, forensics, investigations, and governance (White Paper, January 2024)

Interpol, Technology Assessment Report on Metaverse (Report, 2022)

Jhaver S, Birman I, Gilbert E and Bruckman A, 'Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator' (2019) 26(5) ACM Transactions on Computer-Human Interaction 1, 35

Kasiyanto S and Kilinc M R, 'The Legal Conundrums of the Metaverse' (2022) 1(2) Journal of Central Banking Law and Institutions 299, 322

Keller D, 'The DSA's Industrial Model for Content Moderation' (Verfassungsblog: On Matters Constitutional, 2022). doi:10.17176/20221201-154204-0



Konig P, 'Islamic State group's experiments with the decentralized web' (Europol, 9 April 2019) <<https://www.europol.europa.eu/publications-events/publications/islamic-state-group-s-experiments-decentralised-web>>

Lemmer Webber C, 'Mastodon launches their ActivityPub support, and a new CR!' (ActivityPub, 10 September 2017), <<https://activitypub.rocks/news/2017-09-10-mastodon-launches-their-activitypub-support-and-a-new-cr/>>

Masnick M, 'Protocols, not platforms: A technological approach to free speech' (Knight First Amendment Institute, 21 August 2019) <<https://perma.cc/MBR2-BDNE>> accessed 20 May 2024

Miedico M, 'The Application of Augmented Reality and Virtual Reality Technologies in Countering Terrorism and Preventing Violent Extremism', (United Nations Office for Counter Terrorism, 8 July 2021)

Miller G, 'The Digital Services Act Is Fully In Effect, But Many Questions Remain' (TechPolicy.press, 20 February 2024) <<https://www.techpolicy.press/the-digital-services-act-in-full-effect-questions-remain/>> accessed 30 May 2024

Morgese G, 'Moderazione e rimozione dei contenuti illegali online nel diritto dell'UE' (Federalismi.it, 12 January 2022) <<https://www.federalismi.it/nv14/articolo-documento.cfm?Artid=48110>>

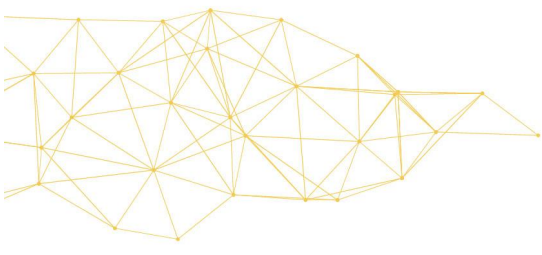
Mosley T R, 'How an undercover content moderator polices the Metaverse' (MIT Technology Review, 28 April 2023) <<https://www.technologyreview.com/2023/04/28/1072393/undercover-content-moderator-polices-the-metaverse/>>

Parsons T D, Courtney C, Cosand L, Iyer A, Rizzo A and Oie K, 'Assessment of Psychophysiological Differences of West Point Cadets and Civilian Controls Immersed within a Virtual Environment' in Schmorow D D, Estabrooke I V and Grootjen M (eds.), Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience (Lecture Notes in Computer Science, Berlin, Heidelberg 514, 23, 2009)

Robertson A, 'How the Biggest Decentralized Social Network Is Dealing with Its Nazi Problem' (VERGE, 12 July 2019) <<https://www.theverge.com/2019/7/12/20690354/mastodon-decentralized-social-network-nazi-problem-gab-mitigation>> accessed 24 May 2024

Rozenshtein A Z, 'Moderating the Fediverse: Content Moderation on Distributed Social Media' (2022) 3 Journal of Free Speech Law 217, Minnesota Legal Studies Research Paper No. 23-19

Shilina S, 'The future of social networking: Decentralization for user empowerment, privacy, and freedom from censorship' (Medium, 6 November 2023) <<https://medium.com/paradigm-research/the-future-of-social-networking->



decentralization-for-user-empowerment-privacy-and-freedom-from-a0a8f74790cb>
accessed 3 May 2024

Teo M, 'The Wellbeing of Content Moderators in the Metaverse: Navigating New Realities' (ZevoHealth, 21 March 2024) <<https://www.zevohealth.com/blog/the-wellbeing-of-content-moderators-in-the-metaverse-navigating-new-realities/>>

Turillazzi A, Taddeo M, Floridi L and Casolari F, 'The digital services act: an analysis of its ethical, legal, and social implications(2023) 15(1) Law, Innovation and Technology 83, 106

van den Branden B, Davidse S, Smit E, 'In between illegal and harmful: a look at the Community Guidelines and Terms of Use of online platforms in the light of the DSA proposal and the fundamental right to freedom of expression' (DSA Observatory, 2 August 2021)<<https://dsa-observatory.eu/2021/08/02/in-between-illegal-and-harmful-a-look-at-the-community-guidelines-and-terms-of-use-of-online-platforms-in-the-light-of-the-dsa-proposal-and-the-fundamental-right-to-freedom-of-expression-part-1-of-3/>>
accessed 28 June 2024

Vogelezang F, 'Illegal vs Harmful Online Content. Some reflections on the upcoming Digital Services Act package' (Institute for Internet and the Just Society, 2 December 2020) <<https://www.internetjustsociety.org/illegal-vs-harmful-online-content>> accessed 28 June 2024

Wagner K, 'Lessons From the Catastrophic Failure of the Metaverse' (The Nation, 3 July 2023) <<https://www.thenation.com/article/culture/metaverse-zuckerberg-pr-hype/>>
accessed 29 May 2024

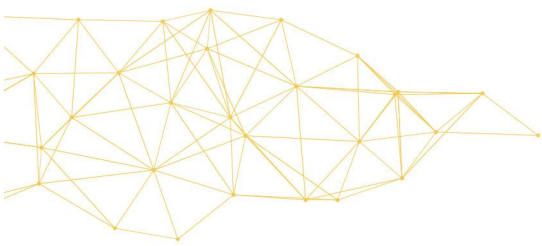
Werbach K., The blockchain and the new architecture of trust (Cambridge, Massachusetts: The MIT Press 2018)

Zignani M, Quadri C, Gaito S, Cherifi H, Rossi G P, 'The Footprints of a "Mastodon": How a Decentralized Architecture Influences Online Social Relationships' (2019) IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) 472, 477

Zulli D, Liu M and Gell R, 'Rethinking the "social" in "social media": insights into topology, abstraction, and scale on the Mastodon social network' (2020) 22(7) New Media & Society 1188, 1205

AA.VV, 'Is decentralized internet the future?' (Security Senses, 17 April 2024) <<https://securitysenses.com/posts/decentralized-internet-future>> accessed 30 April 2024

AA.VV., 'Federated Architecture: learn the Key benefits and how it is different from other architectures' (Atlan, 31 August 2023) <<https://atlan.com/federated-architecture/#what-is-a-federated-architecture>>



AA.VV., '10 Best Metaverse platforms to watch out in 2023' (Blockchain Council, 27 September 2023) <<https://www.blockchain-council.org/metaverse/10-best-metaverse-platforms/>>

AA.VV., 'What is a federated network' (VMWare)
<https://www.vmware.com/topics/glossary/content/federated-network.html>



MetaverseUA
Chair



Universitat d'Alacant
Universidad de Alicante

