CPSL
Content Policy & Society Lab

CONTENT GOVERNANCE IN THE METAVERSE: A PRIMER

AUTHORS:
Julie Owono, Leïla Morch
EDITED BY:
Meri Baghdasaryan
DESIGN BY:
Ombretta Minkue Meye

AUGUST 2022

CONTENTS

About CPSL

The Content Policy and Society Lab, a project funded by the Program on Democracy and the Internet at Stanford University, and the Hewlett Foundation, prototyped a framework for the use of the multistakeholder approach to adress content governance Challenges.

Authors

Julie Owono

Executive Director of the Content Policy & Society Lab (CPSL) and a fellow of the Program on Democracy and the Internet (PDI) at Stanford University. She is also the Executive Director of digital rights organization Internet Sans Frontières, one of the inaugural members of the Meta Oversight Board, and an affiliate at the Berkman Klein Center at Harvard University. She holds a Master's degree in International Law from la Sorbonne University in Paris, and practiced as a lawyer at the Paris Bar.

Leïla Mörch

works in the fields of Digital Governance, Content policy and Web3. She graduated from La Sorbonne, l'Ecole Normale Supérieure, and l'Institut des Mines de Paris. She is the Research Project Coordinator at CPSL.

Edited by

Meri Baghdasaryan

Attorney from Armenia, working on human rights issues at the intersection of law and technology. She is a Steering Committee Member of the IGF Dynamic Coalition on Internet Rights and Principles and an Internet Law and Policy Foundry Fellow. She is the Special Publications Lead at CPSL.

Designed by:

Ombretta Minkue Meye

Digital Content Creator, she holds a Master in Communications and Media technologies from Academy of Art University in San Francisco. She is the Communication assistant at CPSL.

INTRODUCTION

Content governance is complex, and it may face new and interesting challenges as the virtual immersive technologies become more widely used. The current content governance systems are far from perfect, and continuously face new unprecedented situations, such as the COVID-19 pandemic.

As the technological advancements have so far outpaced the policy and regulatory discussions and initiatives, it is high time to start discussing the issues of content governance in the virtual immersive spaces.

The goal of adding a robust content governance system is to ensure safety, security and trust in the online community. In this article, we at Content Policy and Society Lab at Stanford University (CPSL) set the stage for discussions on the particular content governance challenges the metaverse presents.

We will start with a brief overview of the current content governance challenges on online platforms. Then we will discuss the peculiar challenges of content moderation in the metaverse. We will conclude with CPSL's vision on content governance in the metaverse.

I. Into the Metaverse

"While rock climbing with friends, I had a feeling of vertigo." This statement illustrates a different scenario when it refers to a sporting experience in a metaverse. In this context, a metaverse is an immersive reality, involving the use of virtual reality headsets and other equipment. This is different from the so-called "flat" metaverse, such as Decentraland[1] or Second Life.[2] Louis B. Rosenberg,[3] CEO & Chief Scientist of Anonymous AI, defines the metaverse as "the societal transition between "flat" media seen in the third person and immersive media experienced in the first person", which allows us to "go from an outside observer to an inside participant."[4]

In an Interview with CPSL, Nicolas Dupain, President of Immersive Learning Lab,[5] noted that this type of metaverse was part of the Milgram Continuum,[6] illustrated below:

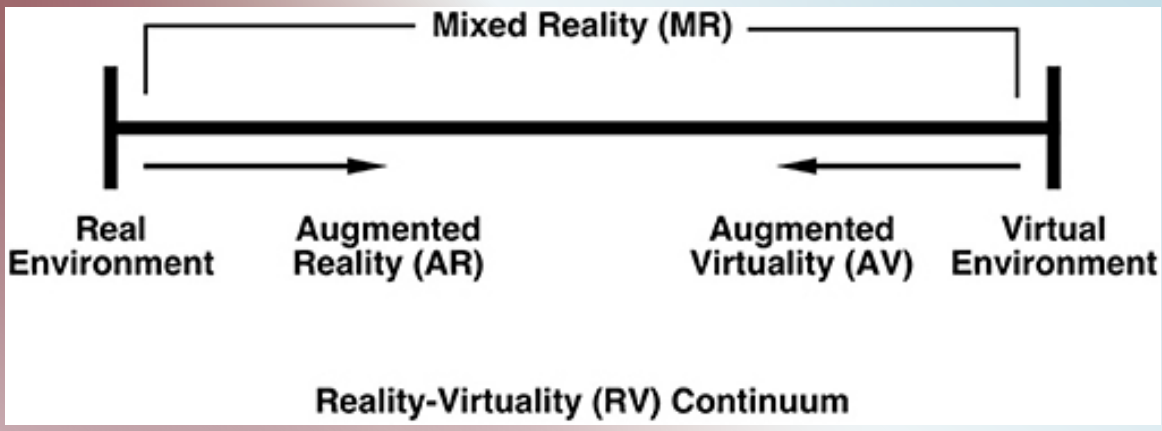

Figure: Milgram and Kishino reality-virtuality continuum (adapted from Milgram et al., 1994).

Hence, in this context, we can define a metaverse[7] as a three-dimensional universe, a collective (unlike a simple universe in immersive reality), accessible via a virtual or augmented reality device, creating an alternative digital reality of immersive content.

1 See https://decentraland.org .
2 See https://secondlife.com/ .
3 Louis Rosenberg, Unanimous AI (2019), available at https://unanimous.ai/staff-item/louis-rosenberg-phd/.
4 See Louis Rosenberg, Metaverse is "societal transition" (Jun. 13, 2022), available at https://www.facebook.com/watch/?v=762980381402201.
5 See Accueil - Immersive Learning Lab, available at https://franceimmersivelearning.fr/immersivele.
6 Milgram, Takemura, Utsumi & Kishino, Augmented reality: a class of displays on the reality-virtuality continuum (Dec. 21, 1995), available at https://www.spiedigitallibrary.org/conference-proceedings-of-spie/2351/1/Augmented-reality--a-class-of-displays-on-the-reality/10.1117/12.197321.short?SSO=1.
7 See e.g., Adrien Ma, What is the metaverse, and what can we do there?, The Conversation (May 23, 2022) available at https://theconversation.com/what-is-the-metaverse-and-what-can-we-do-there-179200.

II. Status quo: content governance systems on social media, and their shortcomings

As our lives continuously grow into the digital domain, from elections and working remotely, to access to information, it is even more crucial to keep online spaces safe and secure. Transparency and international consensus around minimum standards for content policies based on international human rights standards should be among the basic principles of content governance.
CPSL views the governance of online content ont three levels :

- content moderation: consists of sorting the user content on online platforms. It can take several forms, such as algorithmic, manual, upstream or downstream, and it can span from takedown and labeling to downranking and so forth.

- content regulation: set of rules adopted by a legislator or regulator, creating a framework for content moderation.
- content policy: the actual rules that represent a company's societal vision of what online content should be allowed. We, at the Content Policy & Society Lab, believe that content policy should be based on principles and values of international human rights law.

- content governance: a system of norms governing the process of creation and enforcement of content policies, as well as the distribution of powers among entitities responsible for each task. These norms oversee how content is made available online.

In the United States, online service providers are shielded from some liability for third party content under the Section 230 of the Communications Decency Act, though this immunity does not extend to immunity violations of federal criminal law,[8] intellectual property law or electronic communications privacy law. Section 230 incentivizes the online platforms to engage in voluntary content moderation, leaving them with the freedom to shape their own content moderation mechanisms. The extent and depth to which online platforms engage in voluntary online content moderation has attracted a lot of debate in the US,[9] and several bills[10] have been introduced in the Congress, such as the Digital Services Oversight and Safety Act,[11] or Platform Accountability and Transparency Act.[12] These bills are allegedly aimed at providing more user control and enhancing the transparency of the online platforms, at the same time, attempting to reform or repeal the intermediary liability regime under Section 230.

In Europe, the new Digital Services Act (DSA) sets out the obligations of online platforms to limit the spread of illegal content online.[13] However, it does not define what constitutes illegal content, leaving it to the European Union's (EU) Member States. This means that the online platforms should, first, comply with national laws, and second, go beyond the national regulations, adopting their own rules For instance, during the COVID-19 pandemic, Twitter banned dissemination of false information about the disease that could cause harm.[14] Facebook has alsoentered into this war against the infodemic by opening a part of its advertising program free of charge to public health organizations like the World Health Organization and strengthening its partnerships around fact-checking.[15]

8 See e.g., Eric Goldman, An Overview of the United States' Section 230 Internet Immunity (Jan. 9, 2019), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3306737.
9 See e.g., David Morar & Chris Riley, A guide for conceptualizing the debate over Section 230, Brookings (Apr. 9, 2021), available at https://www.brookings.edu/techstream/a-guide-for-conceptualizing-the-debate-over-section-230/.
10 See more Section 230 Legislation tracker, available at https://citrispolicylab.org/section230/.
11 H.R.6796 - 117th Congress.
12 S.797 - 117th Congress.
13 See e.g., Access Now, The Digital Services Act: Your Guide to EU's New Content Moderation Rules (Jul. 6, 2022), available at https://www.accessnow.org/digital-services-act-eu-content-moderation-rules-guide/.
14 Twitter Help Center, How we Address Misinformation on Twitter, https://help.twitter.com/en/resources/addressing-misleading-info.
15 See Mark Zuckerberg, Facebook post (Mar. 3, 2020), available at https://www.facebook.com/4/posts/10111615249124441/?d=n.

However, despite these regulatory advances, content moderation rules are still not well equipped to withstand the challenges of our time. The conflict in Ukraine is a clear proof of this: in the absence of an international framework and transparent and balanced rules to deal with the information war waging on the online platforms.[16] Another example is the fight against online disinformation. Despite recent reports[17] or efforts, particularly at the European level,[18] the Ethiopian[19] and Ukrainian [20] crises show the impact that the disinformation can have on global geopolitics and the real violence on the ground that it can lead to.

In a Le Monde op-ed,[21] CPSL underlined the need for a more advanced,realistic and far-sighted regulatory framework, to put an end to the arbitrary power of platforms. We highlighted that access to data should not be treate as the only measure for compliance with regulatory requirements, and noted the importance of understanding the global societal impact of online platforms.

Content governance on online platforms is currently insufficient for several reasons:

First, the governance frameworks suffer from a lack of resources: there are not enough human moderators,[22] that would be able to handle the review of online content at scale.[23]

Next, content moderation is context specific, and ensuring consistency globally for decisions involving different cultural and politicalenvironments is a difficult task.[24] Most of the so-called Big Techcompanies emerged in countries with common law traditions,[25] which can be in contrast to understandings, e.g., around free expression andprivacy in othe corners of the world.[26] In addition, global content decisions may suffer from lack of understanding of local context,[27] and experience slower reactions from platforms.[28] This disproportionately affects communities outside the Global North, especially having adverse implications on marginalized groups.[29]

16 See e.g., Mark Scott, As War in Ukraine Evolves, So Do Disinformation Tactics, Politico (Mar. 10, 2022), available at https://www.politico.eu/article/ukraine-russia-disinformation-propaganda/; Sara Brown, In Russia-Ukraine War, Social Media Stokes Ingenuity, Disinformation, MIT Sloan Blog (Ap. 6, 2022), available at https://mitsloan.mit.edu/ideas-made-to-matter/russia-ukraine-war-social-media-stokes-ingenuity-disinformation.

17 See e.g., Rachel Griffin, The Bronner report on disinformation hints at Macron's political priorities, Tech Policy Press (Jan. 27, 2022), available at https://techpolicy.press/the-bronner-report-on-disinformation-hints-at-macrons-political-priorities/.

18 2022 Strengthened Code of Practice on Disinformation, European Commission, January 16, 2022, available at https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

19 See e.g., The Documentary Podcast : Ethiopia's Disinformation War, Apple Podcasts.

20 See e.g., Molly Killeen, Ukraine disinformation years in the making, say lawmakers, EURACTIV, (Mar. 8, 2022) available at https://www.euractiv.com/section/digital/news/ukraine-disinformation-years-in-the-making-say-lawmakers/.

21 CPSL, et. al, 2022 Must be the year of open algorithms, Le Monde (Feb. 8, 2022), available at https://www.lemonde.fr/idees/article/2022/02/08/2022-doit-etre-l-annee-de-l-ouverture-des-algorithmes_6112825_3232.html.

22 See e.g., Mark Scott & Laura Kayali, What Happened When Humans Stopped Managing Social Media Content, Politico (Oct. 21, 2020), available at https://www.politico.eu/article/facebook-content-moderation-automation/

23 See e.g., Sam Levin, Google to hire thousands of moderators after outcry over YouTube abuse videos, The Guardian (Dec. 5, 2017), available at https://www.theguardian.com/technology/2017/dec/04/google-youtube-hire-moderators-child-abuse-videos.

24 See e.g., Robyn Caplan, Content or Text Moderation, Data and Society (2018), available at https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf.

25 See e.g., Tom Simonite, Facebook Is Everywhere; Its Moderation Is Nowhere Close, WIRED (Oct. 25, 2021), available at https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/; Delia Marinescu, Facebook's Content Moderation Language Barrier, New America, (Sep. 8, 2021), available at https://www.newamerica.org/the-thread/facebooks-content-moderation-language-barrier.

26 See e.g., Marwa Fatafta, Facebook is bad at moderating in English. In Arabic, it's a disaster, Rest of World (Nov. 18, 2021), available at https://restofworld.org/2021/facebook-is-bad-at-moderating-in-english-in-arabic-its-a-disaster/.

27 See e.g., Dan Milmo, Rohingya sue Facebook for £150bn over Myanmar genocide, The Guardian (Dec. 6, 2021), available at https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence.

28 See e.g., Kate Ly Johnston, Young Vietnamese Americans Say Their Parents Are Falling Prey To Conspiracy Videos, BuzzFeed News (Apr. 21, 2021), available at https://www.buzzfeednews.com/article/katejohnston2/vietnamese-american-youtube-misinformation-covid-vaccine.

29 See e.g., Robyn Caplan, Content or Text Moderation, Data and Society (2018), available at https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf.

Moreover, automated content moderation systems are prone to mistakes and may disproportionately limit human rights.[30] It is challenging to analyze the context of human speech across different language and cultural boundaries. Moreover, the transparency and accountability of the content moderation algorithms has been widely criticized,[31] and is the subject of an upcoming EU regulation on Artificial Intelligence.[32]

Finally, the lack of consensus around definitions of, e.g., hate speech, disinformation, or terrorist organizations, leads to differing practical treatments of similar situations. Many association fighting against online hate emphasize the persistence of the phenomenon,[33] despite the platforms'moderation efforts. In the context of dangerous organizations and violent extremism, theTaliban, banned from Facebook that considers them aterrorist group, is nevertheless accepted on Twitter, which serves as a platform for their propaganda.[34]

In the case of the conflict in Ethopia, the Ethopian Parliament designated two groups, Tigray People's Liberation Front (TPLF) and the Oromo Liberation Army (OLA), as terrorist organizations, while Meta only listed OLA as such, raising concerns of biased moderation.[35] These decisions, taken unilaterally by private platforms, sometimes without justification, and often without coordination between major platforms, have significant political, geopolitical, and societal repercussions. However, since the escalation of the conflict in Ukraine in February 2022, we have witnessed more coordinated and timely responses from online platforms.[36]

30 See e.g., Spandana Singh, Everything in Moderation, New America (Last Updated: Jul. 22, 2019) available at https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/the-limitations-of-automated-tools-in-content-moderation/.

31 See e.g., Robyn Caplan, Joan Donovan, Lauren Hanson & Jeanna Matthews, Algorithmic Accountability:A Primer, Data & Society (Apr. 18, 2018), available at https://datasociety.net/library/algorithmic-accountability-a-primer/.

32 See e.g., Alex Engler, The EU AI Act Will Have Global Impact, But a Limited Brussels Effect, Brookings (Jun. 8, 2022), available at https://www.brookings.edu/research/the-eu-ai-act-will-have-global-impact-but-a-limited-brussels-effect/.

33 See e.g., Online hate: associations ask for an expertise in court on TwitterEuropean Observatory of Online Hate (May 26, 2021), available at https://www.europe1.fr/technologies/haine-en-ligne-des-associations-demandent-en-justice-une-expertise-chez-twitter-4047702.

34 See e.g., Thomas H Johnson, Matthew DuPee, Wali Shaaker, The Taliban's Use of the Internet, Social Media Video, Radio Stations, and Graffiti (2018), available at https://www.researchgate.net/publication/331220278_The_Taliban's_Use_of_the_Internet_Social_Media_Video_Radio_Stations_and_Graffiti.

35 See e.g., Robera Hamda, Facebook's 'Dangerous Individuals and Organizations' list concerns Oromo users in Ethiopia,Global Voices (Oct. 28, 2021), available at https://globalvoices.org/2021/10/28/facebooks-dangerous-individuals-and-organizations-list-concerns-oromo-users-in-ethiopia/.

36 See e.g., Rebecca Kern, Mark Scott, Social media platforms on the defensive as Russian-based disinformation about Ukraine spreads, Politico (Feb. 24, 2022), available at https://www.politico.com/news/2022/02/24/social-media-platforms-russia-ukraine-disinformation-00011559.

III. Challenges of Content Governance in the Metaverse

The overview above clearly shows that current online content governance frameworks are not well equipped to withstand the scale and amount of current online content. The metaverse is a source of additional challenges compared to social media. In this section, we will walk through some concerns in the context of content governance in the metaverse.

Similar to other digital technologies, the development of the metaverse outpaces the adoption of regulatory responses. And it is high time to draw inspiration from the lessons learnt from social media online content moderation and build digital spaces, based on principles of international human rights law.

a. Spotlight on Privacy

Allowing better access to data, enabling the researchers to understand its effects and their interactions with algorithms, is anessential step in building an effective governance framework, especially for viral content.

The arrival of the metaverse and virtual reality headsets poses many challenges. The EU's General Data Protection Regulation (GDPR) stipulates special protections for sensitive data, including, personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership.[37] In a recent ruling, the Court of Justice of EU held that processing personal data liable to indirectly disclose sexual orientation constitutes processing of special categories of personal data.[38]

While various European Data Protection Authorities continue to report data breaches,[39] Metaverses may entail data collection on a new level. This may include collection of even more sensitive data, ranging from biometric information to potential emotion detection. For example, French startup Next Mind offers equipment to monitor brain activity live: the developed a lightweight neural interface technology that measures brain's electrical activity when the user 'clicks' something with their eyes.[40] The risks of biometric data breaches[41] can in many ways be more dangerous than hacking shopping preferences. Thus, privacy and data protection should be front and center, as metaverses are in design and development.

37 Article 9, Regulation (EU) 2016/679, available at https://gdpr-info.eu/art-9-gdpr/.
38 See e.g., Natasha Lomas, Sensitive data ruling by Europe's top court could force broad privacy reboot, TechCrunch (Aug. 2, 2022), available at https://techcrunch.com/2022/08/02/cjeu-sensitive-data-case/.
39 See e.g., CNIL, Use of Google Analytics and data transfers to the United States: the CNIL orders a website manager/operator to comply (Feb. 10, 2022), available at https://www.cnil.fr/en/use-google-analytics-and-data-transfers-united-states-cnil-orders-website-manageroperator-comply.
40 See e.g., Brian Heater, Snap Buys Mind-Controlled Headband Maker NextMind, TechCrunch (Mar. 23, 2022), available https://techcrunch.com/2022/03/23/snap-buys-mind-controlled-headband-maker-nextmind/.
41 See e.g., Steve Symanovich, Biometric data breach: Database exposes fingerprints and facial recognition data of 1 million people (Aug. 18, 2019), available at https://us.norton.com/internetsecurity-emerging-threats-biometric-data-breach-database-exposes-fingerprints-and-facial-recognition-data.html#.

This protection can even go towards commitments not to collect data. In late 2021, Meta announced that it would end the use of facial recognition on the photos of Facebook users.[42] This type of commitment could also be adopted with respect to data collected in virtual worlds. According to Meta's internal documents,[43] the company does not know where a lot of collected user data goes, making it difficult to comply with regulations on data protection in different jurisdictions.

While Meta is not an isolated example, the strengthening of data protection is paramount. Moreover, the relationships between different metaverses are also unclear. Whether different approaches and perspectives on content policies from online platforms would bring about different metaverses, such as those with Taliban members or without, remains to be seen.

b. Content Moderation in Virtual Worlds

Ultimately, the impact of content moderation is felt on individual levels, and collectively, on the societal level. It is interesting to compare with those made on traditional social media like Twitter, Facebook, YouTube etc., since in many aspects, the metaverse is an "augmented" social media.

A 2019 Meta Reality Lab[44] study found that efforts to regulate harassment in the metaverse would be challenging. Such features as synchronous voice chat, heightened feelings of presence and embodiment, and avatar movements may contribute to the sense of harassment in the virtual space. Moreover, "the distributed landscapes of virtual reality applications and the dynamic nature of local community norms" contribute to the challenges.[45] **Virtual reality spaces seem neither better nor worse than other online spaces. But in virtual worlds the impact of toxic behaviors on users could increase tenfold**.

It would seem that the behavior of people evolves according to their digital representations, also known as the Proteus effect.[46] Observed in several video games, this effect becomes particularly visible in immersive reality settings, where the mind and body cannot differentiate virtual experiences from real ones. **These universes will have a strong impact on our bodies and brains and we must prepare for this. Thinking upstream, the mass use of these technologies will allow us to create effective moderation tools, based on appropriate and transparent rules.**

42 See e.g., Jerome Pesenti, An Update On Our Use of Face Recognition, Meta Newsroom (Nov. 2, 2021), available at https://about.fb.com/news/2021/11/update-on-use-of-face-recognition/.
43 See e.g., Lorenzo Franceschi-Bicchierai, Facebook Data Lineage Internal Document, available at https://www.documentcloud.org/documents/21716382-facebook-data-lineage-internal-document.
44 Lindsay Blackwell, Nicole Ellison, Natasha Elliot-Deflo & Raz Schwartz, Harassment in Social Virtual Reality: Challenges for Platform Governance (Nov. 7, 2019), available at https://dl.acm.org/doi/abs/10.1145/3359202.
45 Id.
46 See e.g., Nick Yee & Jeremy Bailenson, The Proteus Effect: The Effect of Transformed Self-Representation on Behavior, available at https://stanfordvr.com/mm/2007/yee-proteus-effect.pdf.

In addition to posing ethical, normative and conceptual challenges, and the increased impact of collective immersive reality spaces on users, moderation would encounter new technical challenges. One of the major challenges will be that of real-time moderation. Unlike social media where publications are stored on a medium, in the metaverse the exchanges will be visual and audio. These do not always seem to be stored as is the case on the VR chat platform. Their terms of use do not seem to address this issue,[47] but the company seems to advocate for active user documentation of violations of community standards.[48] Horizon, Meta's immersive social media, has developed a system enabling the platform moderators to re-watch the few minutes of an incident, when deciding on a user report.[49]

Monitoring billions of interactions in real time will be a challenge for platforms and will require very significant costs. In a March 2021 internal memo to Meta teams,[50] Andrew Bosworth, developer at Meta, in charge of virtual and augmented reality, noted that the nature of the immersive virtual realities may exacerbate the harassment in these spaces, but added that the theory of change is to evolve the culture to have less enforcement action.[51]

Hence, the ethical, normative and conceptual challenges require not only thinking about tools and moderation frameworks inspired by current rules and methods but also taking into account the technological constraints brought by the immersive and real-time nature of these universes, as well as the feeling of embodiment given to players or users.

47 VRChat, Terms of Service (Last Updated: Aug. 10, 2022), available at https://hello.vrchat.com/legal.
48 See If you visit public worlds often and have the capability to record your sessions, r/VRchat, available at https://www.reddit.com/r/VRchat/comments/bnu4dv/discussion_if_you_visit_public_worlds_often_and/.
49 Horizon Safety Video (Oct. 12, 2020), available https://www.youtube.com/watch?v=XpfEw65X7F0.
50 Hannah Murphy, How will Facebook keep its metaverse safe for users?, Financial Times (Nov. 12, 2021), available at https://www.ft.com/content/d72145b7-5e44-446a-819c-51d67c5471cf/.
51 Id

CASE STUDY:
The use of a multistakeholder approach to develop content rules for the VRCHAT communities

In order to better understand the challenges of creating a content governance system in the metaverse, we collected testimonies from users and moderators of these spaces. We were able to talk to two daily users of the VRChat platform. It is a “social” virtual reality platform, but the platform is defined as a video game, proof of the thin line between the two universes.

For confidentiality reasons we cannot reveal their identity or their online pseudonym. One of them (we will refer to him as Robin), a user/influencer of the metaverse, in particular of VR chat, seeks to popularize the use of this platform and explain and demonstrate the metaverse in virtual reality. The second interviewee (we will refer to him as Arnaud), is a volunteer moderator for the French-speaking community of VRchat. Both expressed concerns over the lack of moderation tools from the platform, especially for content in French.

According to Arnaud, despite the low number of users on the VRChat platform (compared to traditional social media), moderation is, so far, failing. Since 2018, he has been able to see a degradation of the French-speaking moderation from the platform’s professional teams, leading to a need for moderation by users. However, the users have limited tools: the main sanction against a user acting in a harmful way on the metaverse space of the French VR community is the ejection of the user from the platform for 1 hour. If a more severe sanction is needed, a moderator from the VRchat teams should be involved. The major issue is then the ability to keep track of the user’s bad behavior. The ease of changing a nickname is the first obstacle. At the same time, it is easy to create a new account using a different IP address.

An automated, privacy-centric and blockchain-based system to verify proof of identity at the time of registration could be one solution. User communities are already discussing these issues, as they will first face the consequences of the platforms’ technological choices.

IV. Towards a democratic content governance system in the metaverse: CPSL's approach

As illustrated above, the status quo in content governance would be insufficient to address the moderation challenges in virtual immersive worlds. The development of the metaverse is an opportunity to learn from the past, to build a next iteration of the web that is safe, secure, inclusive, and beneficial to all.

CPSL's proposal is that the metaverse be based on rules that guarantee safety in these immersive public spaces. These rules should be pre-established and based on rule of law principles and values, which are the only way to guarantee the legitimacy of these rules.

The CPSL is exploring what this system of democratic governance of online content might look like, using the multi-stakeholder method. It can be defined as the approach by which "*people and organizations from different domains participate alongside each other in sharing ideas or developing common policy.*"[52] The multistakeholder approach is a way of thinking about how to achieve this. It is recognized as particularly appropriate given the nature of the Internet.[53]

The CPSL is studying the effects of using this approach, from the creation of content rules to their enforcement, and is interested in their potential to foster content regulation, whether state or private, that is effective, legitimate, and based on international human rights standards.

Dialogue and collaboration between governments, diplomats, researchers, companies, civil society representatives, content moderators, and users contribute to a better understanding of the challenges faced by each respective actor, and allow them to imagine solutions anchored in the search for consensus.

These dialogues and collaborations between actors lead to the emergence of common points on the main characteristics of an online content governance system.

52 See e.g., Internet Society, Why the Multistakeholder Approach Works (Apr. 26, 2016), available at https://www.internetsociety.org/resources/doc/2016/internet-governance-why-the-multistakeholder-approach-works/.
53 Id.

A. An interoperable system

Mark Zuckerberg, founder of Meta, advocates[54] for a technically interoperable metaverse, where the avatars and digital objects owned by users accompany them in different worlds. This is also the purpose of the Metaverse Standards Forum, which is aimed at creating a space where "leading standards organizations and companies cooperate to promote interoperability standards for an open metaverse."[55] The Forum will not set standards, but will serve as a coordination center for efforts and resources in this space. The entity already brings together industry representatives, such as Microsoft, Meta, Huawei, Ready Player Me, Samsung, Adobe, Nvidia, and Epic Games.[56]

No such discussion yet exists on content rules, particularly those rules aimed at ensuring public safety in the various metaverses. Take the example of the Taliban designation discussed above. It is unclear how different meraverses will treat such organizations, that can be viewed as terrorist, but can also become a de facto government entity. The questions of rule adoption, their format, and enforcement oversight remain open.

B. An inclusive system

Another key proposal is based on the model of Middleware,[57] or decentralized moderation.[58] On the one hand, companies set rules for their platforms that contain high level language, and may be ambiguous to the users. On the other hand, there is a transfer of the responsibility of maintaining order in private spaces to moderators who are also users. This is understandable, as it is important to maintain private spaces, even in the metaverse. However, such a transfer raises many questions, such as moderator support, appeals process and the role moderators play in setting the rules.

54 Mark Zuckerberg, Facebook Post (Jun. 22, 2022), available at https://www.facebook.com/zuck/posts/pfbid021Q1ZMJFWcXBvPcaqpLWgp3x5HTMD3vyVNbPB17BcuX6n1UmAKh7Kyv6dUb6oKP7Nl
55 The Metaverse Standards Forum, available at https://metaverse-standards.org/.
56 Id.
57 See e.g., Daphne Keller, The Future of Platform Power: Making Middleware Work, 32 Journal of Democracy 168 (2021), available at https://www.journalofdemocracy.org/articles/the-future-of-platform-power-making-middleware-work/.
58 See e.g., David A. M. Goldberg, Claudia Lo & Margeigh Novotny, How Content Moderation and Anti-Vandalism Works on Wikipedia (Jul. 30, 2020), available at https://design.wikimedia.org/blog/2020/07/30/content-moderation-anti-vandalism-wikipedia.html.

C. A transparent system

There is no question that so-called Big Tech companies wield increasing power, on discourses, opinions, behaviors, that are considered acceptable online, and therefore in our societies. This de facto situation is rightly criticized from a democratic perspective,[59] and some proposals attempt to find ways to create equivalents of checks and balances mechanisms that would limit this power.Transparency, particularly meaningful transparency,[60] has been in recent years pointed out as one possible way to protect users' rights in the technology industry.[61] This is for instance the case for the right to effective remedy: referring to a recommendation by the UN Special Rapporteur on Freedom of Expression related to business entities' responsibility to provide remedy to adverse Human Rights practices, the Meta Oversight Board urged the company to expand its transparency efforts to ensure that users are aware of the use of automation for moderation of their content.[62]

D. A system accountable for its errors

Transparent content governance mechanisms of companies operating in the metaverse allow opportunities for accountability. The latter is necessary to create, or restore, trust and legitimacy of online spaces that promise to incarnate a virtual version of our analogic lives, and will eventually continue to hold troves of information about us.

E. The importance of proactive state regulations

The most recent content regulations, notably the DSA,[63] seem to have a limited grasp on immersive social media. In it's op-ed published in Le Monde on algorithmic transparency, the CPSL questioned whether the DSA's imposed systemic risk assessment would apply only to ensure companies' compliance with the new legislation obligations, or could potentially also have prospective applications for research and development, including for products and services that don't exist yet. It is also worth mentioning that even if the GDPR applies to requirements for the transfer of data collected on the virtual worlds outside the EU, the massive amounts of data collected and generated in the metaverse could create some challenges for the applicability of this legislation.

It is important to take stock of past mistakes and failures, and seize the opportunity to proactively design the rules for tomorrow's digital worlds.

59 See e.g., Aviv Ovadya, Towards Platform Democracy: Policymaking Beyond Corporate CEOs and Partisan Pressure, Belfer Center for Science and International Affairs (Oct. 18, 2021), available at https://www.belfercenter.org/publication/towards-platform-democracy-policymaking-beyond-corporate-ceos-and-partisan-pressure.
60 See e.g., Caitlin Vogus & Emma Llansó, Making Transparency Meaningful: A Framework for Policymakers, Center for Democracy and Technology (Dec. 14, 2021), available at https://cdt.org/insights/report-making-transparency-meaningful-a-framework-for-policymakers/.
61 See e.g., Robert Gorwa & Timothy Garton Ash, Democratic Transparency in the Platform Society, Social Media and Democracy (Aug. 24, 2020), available at https://www.cambridge.org/core/books/social-media-and-democracy/democratic-transparency-in-the-platform-society/F4BC23D2109293FB4A8A6196F66D3E41.
62 See e.g., Oversight Board, Case decision 2021-006-IG-UA (Jan. 28, 2021), available at https://www.oversightboard.com/decision/IG-7THR3SI1.
63 See e.g., Daphne Keller, What Does the DSA Say?, The Center for Internet and Society (Apr. 25, 2022), available at http://cyberlaw.stanford.edu/blog/2022/04/what-does-dsa-say-0.

CONCLUSION: The Way Foward

The challenges of content governance in the metaverse put the imperfections of current content governance frameworks under spotlight. On one hand, the metaverse brings along new technical and regulatory challenges. But on the other hand, it provides an opportunity to learn lessons and apply better solutions to content governance issues, allowing us to take the full advantage of technological innovations and enjoy safe, secure, and trustworthy virtual spaces.

As metaverses may have cognitive, emotional and even societal impacts, the platforms that host them must already set the foundations for interoperable, transparent and inclusive virtual spaces. A clear and transparent content governance system may guarantee the legality and legitimacy of the metaverse rules. And a multistakeholder approach to formulating and monitoring content rules is an effective way to create an accountable, inclusive and transparent system.

CPSL envisions a content governance system, that is interoperable and multistakeholder by design, is based on international human rights principles and is equipped with effective enforcement tools.This is an ambitious but essential vision to strive for.