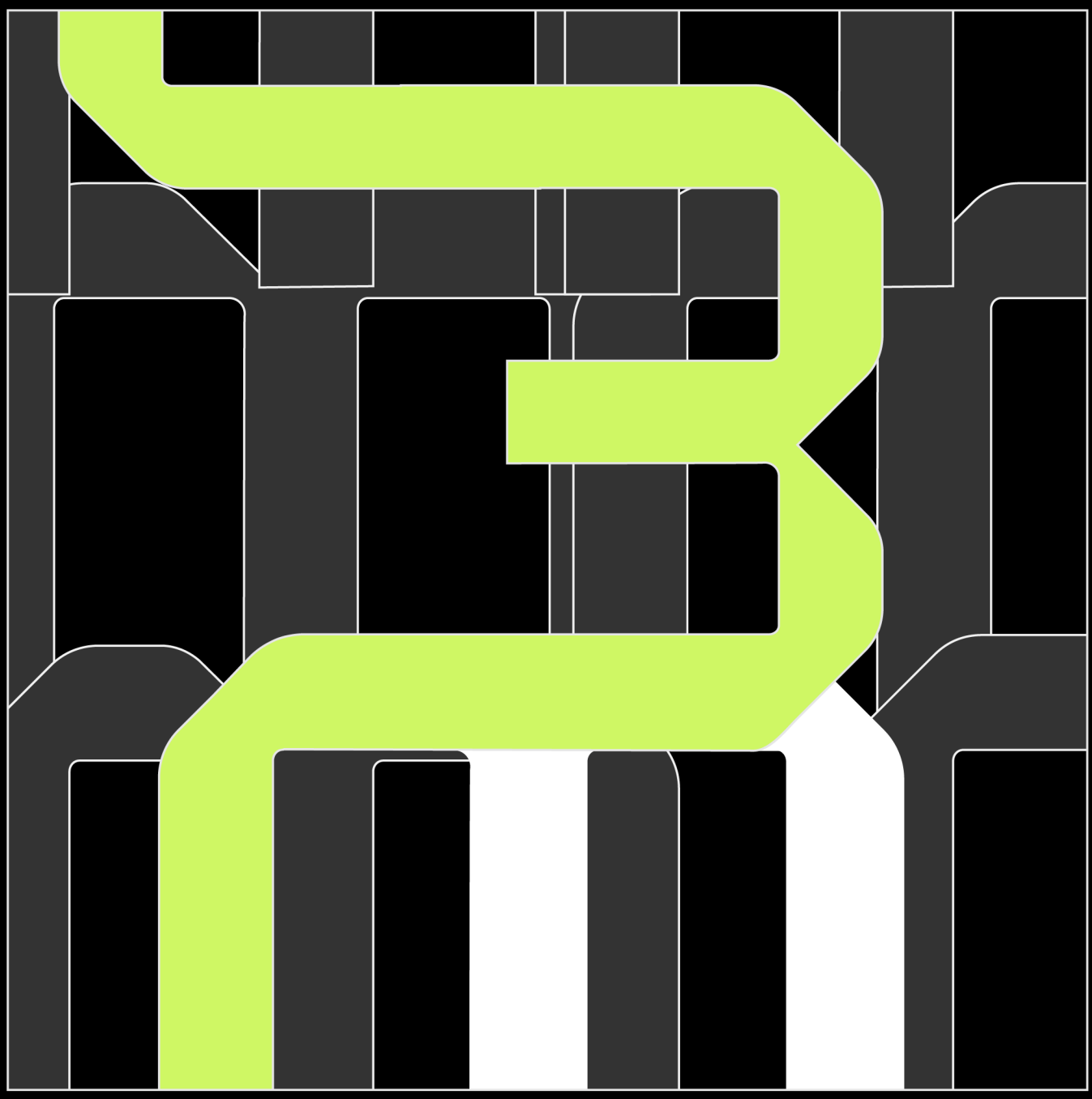
DIVERSO

LABORATÓRIO DE ESTUDOS SOBRE O METAVERSO

DESAFIOS E OPORTUNIDADES DA MODERAÇÃO DE CONTEÚDO NO METAVERSO

Relatório

Artur Pericles Lima Monteiro



Eixo

05

Foco do eixo

MODERAÇÃO 3.0

ITS

RELATÓRIO

Desafios e oportunidades da moderação de conteúdo no metaverso

Autores
Artur Pericles Lima Monteiro

Revisão
João Victor Archegas,
Christian Perrone e
Bernardo Accioli de Vasconcellos

Coordenação
Christian Perrone

Design
Stephanie Lima

Este relatório foi desenvolvido para o diVerso: laboratório de estudos sobre o metaverso do Instituto de Tecnologia e Sociedade do Rio de Janeiro (ITS Rio).

MISSÃO | diVerso

A Ascensão do Metaverso e a Próxima Web: Nossa Missão

O metaverso pode ser definido como a convergência do mundo físico com o mundo digital, consolidando, portanto, um espaço virtual onde as pessoas, interagindo por meio de "avatares", poderão trabalhar, socializar, negociar, jogar e consumir.

O termo foi cunhado pela primeira vez em 1992 no livro de ficção **Snow Crash** de Neal Stephenson, onde avatares realistas interagiam em espaços virtuais tridimensionais. Embora o conceito de metaverso não seja uma novidade, hoje, graças ao desenvolvimento de diferentes tecnologias, já é possível vislumbrar um futuro que será revolucionado por esse novo estágio da era digital.

Ainda assim, segundo as projeções mais otimistas de especialistas e atores centrais do ramo da tecnologia, a exemplo da Meta - empresa que chegou a mudar seu nome em 2021 para refletir sua visão para a Internet do amanhã -, o metaverso só deve se tornar ubíquo em dez anos. Até lá, muitas questões envolvendo os seus impactos na sociedade serão levantadas e debatidas por representantes de diferentes setores. Já é possível vislumbrar alguns sinais dessa dinâmica, como em casos de assédio sexual de mulheres no **Horizon Worlds**, cuja versão beta hoje só pode ser acessada por usuários nos Estados Unidos e Canadá.

O prenúncio de uma tecnologia disruptiva com uma década de antecedência é uma oportunidade ímpar para instituições que, como o ITS Rio, se preocupam em estudar a interseção entre tecnologia e sociedade com o objetivo de abrandar seus impactos negativos e potencializar os

positivos, tudo isso com uma visão voltada aos desafios particulares do Sul Global. Com o objetivo de mobilizar os atores interessados em contribuir com essa empreitada a partir de um ponto de vista multissetorial, o ITS Rio está estruturando o **diVerso**, um **laboratório de estudos sobre o metaverso** com três missões transversais:

a. Fomentar uma comunidade multissetorial de especialistas;
b. Investir na capacitação de seus integrantes (capacity building);
c. Estabelecer uma agenda de discussões e pesquisas na América Latina.

Essas missões vão informar o desenvolvimento de análises e investigações aprofundadas em pelo menos seis eixos verticais:

1. Regulação, Jurisdição e Interoperabilidade;
2. Democracia e Governança;
3. Economia e Propriedade;
4. Gênero, Raça e Proteção de Crianças e Adolescentes;
5. Moderação de Conteúdo e Comportamento;
6. Identidade e Trabalho.

Em cada uma dessas dimensões, os colaboradores do diVerso vão (i) **investigar os impactos da aplicação do metaverso no Sul Global, em especial na América Latina**, (ii) **mapear arranjos regulatórios existentes na região que podem ser reaproveitados**, (iii) **identificar lacunas regulatórias que deverão ser preenchidas por legisladores e outras autoridades públicas e, a partir dos novos desdobramentos tecnológicos**, (iv) **indicar quais são as tendências para o metaverso no futuro**.

Ou seja, trata-se de um esforço conjunto, focado na realidade do Sul Global, que possibilitará o desenvolvimento consciente de um ecossistema regulatório em torno dessa nova tecnologia. Note-se, entretanto, que o objetivo não é precipitar a regulação do metaverso, mas apenas subsidiar o debate que se desdobrará nos próximos anos.

INTRODUÇÃO

Desafios e oportunidades da moderação de conteúdo no metaverso

ARTUR MONTEIRO

Quando falamos em moderação de conteúdo, muito pode estar em jogo: o que pode ser dito em espaços tão diferentes quanto Facebook, Reclame Aqui e TudoGostoso, por exemplo. Não só as regras do que pode ser dito, mas também quem as estabelece e como são aplicadas são questões importantes e a respeito das quais há muita discussão. O metaverso amplifica muitas dessas questões e abre outras. Ao mesmo tempo, também pode oferecer oportunidades para modelos diferentes de governança da esfera digital, menos centralizados em empresas sediadas do Norte Global, como este breve ensaio sugere.

PARTE I

O que é o metaverso?

O que hoje chamamos de metaverso inclui muitos serviços, ambientes e realidades distintas. Uma forma de compreender essa noção enfatiza uma convergência do mundo físico com o mundo digital. Essa convergência pode ser vista como um processo em curso que abarca até mesmo redes sociais tradicionais (diríamos, diante do metaverso) como Facebook, TikTok e Twitter, além de serviços de videoconferência como o Zoom: o que acontece nesses espaços já integra a vida cotidiana, ao mesmo tempo em que não está situado neste ou naquele local geográfico.

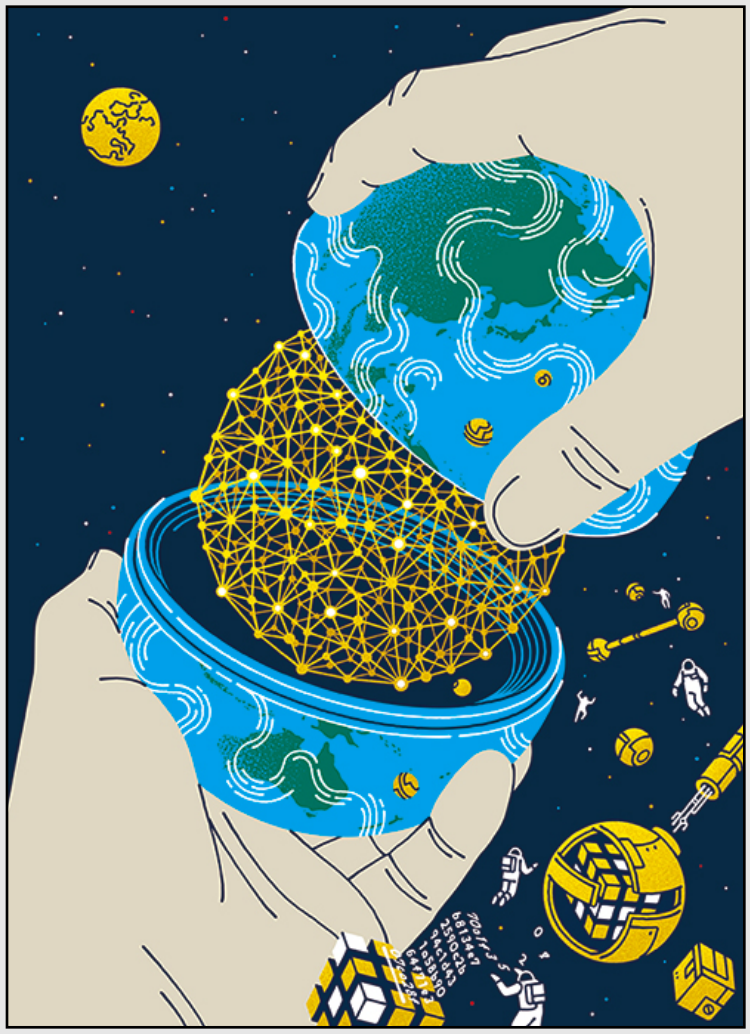


Imagem 1 - Ilustração por Kouzou Sakai.

No entanto, o que em geral está em jogo quando se fala do metaverso é uma forma diferente de convergência, particularmente a partir de dispositivos de *hardware* que proporcionam imersão numa realidade digitalmente sobreposta à realidade que pode ser vista a olho nu, ou num ambiente digital construído que oferece uma outra realidade por completo. No primeiro caso, falamos em realidade aumentada (AR, a partir da expressão em inglês *augmented reality*); no segundo, em realidade virtual (VR, *virtual reality*).

Mesmo a imersão auxiliada por dispositivos como óculos de AR ou VR não é estritamente necessária para o acesso ao metaverso, contudo[1]. O metaverso também é usado para descrever realidades virtuais que não exigem esse tipo de *hardware* especial que pode parecer ainda futurístico e com preços proibitivos para a maior parte da população do Sul Global. Plataformas para videogames como Roblox e jogos como Fortnite sãodiscutidos como pertencentes ao metaverso. E aqui encontramos exemplos bem anteriores à atual onda de interesse no metaverso, como Second Life, lançado em 2003, e *massively multiplayer online roleplaying games* (MMORPGs, algo como jogos de interpretação de personagens online e em massa para multi-jogadores) extremamente populares, como World of Warcraft[2]. Mesmo Roblox (lançado em 2006) e Minecraft (em 2009), bastante citados com o interesse renovado no metaverso, não começaram agora, ainda que tenham passado por etapas de desenvolvimento desde seu lançamento e que tenham alcançado bases de usuários substantivas mais recentemente.

Imagem 2 - Ilustração por Rebecca Mock

Em parte, portanto, o metaverso já está aqui – e há algum tempo. Assim, a experiência com esses precursores pode oferecer aprendizados para as questões que se colocam com a atual onda do metaverso. Isso certamente não quer dizer que não tenhamos questões novas para a moderação de conteúdo. A escala e a abrangência hoje imaginadas para o metaverso criam desafios. Será preciso lidar não só com centenas de milhões de usuários, mas também com perfis bastante diferentes entre esses mesmos usuários (ao contrário, por exemplo, do que poderia ser verdade para MMORPGs). Além disso, muitas propostas pretendem explorar o metaverso em contextos novos: não só para games ou para uma experiência num espaço virtual, mas também reuniões de trabalho, terapia, meditação e *shows*, numa lista que entusiastas dirão não ter fim. Para moderação de conteúdo, isso significa ter de lidar com públicos, usos e contextos diferentes, que podem exigir novas políticas, ferramentas e estruturas de governança.

Imagem 3 - Ilustração por Lena Vargas.

PARTE II

Moderação de conteúdo e desafios no metaverso

A moderação de conteúdo pode abranger tanto conteúdo ilícito, que viola a legislação aplicável (por exemplo, um post que caracteriza crimes de ameaça ou de incitação, como tipificados no Código Penal), quanto conteúdo irregular, que (embora lícito) viola as políticas de conteúdo estabelecidas para aquele ambiente. Essas políticas de conteúdo podem ser apresentadas em documentos que buscam codificar as regras de uma determinada rede social (por exemplo, as Regras do Twitter, ou as Diretrizes da Comunidade do Facebook), em outros documentos publicados pelo provedor, ou até mesmo em posts e páginas específicas a cada comunidade dentro de uma plataforma (como no Reddit).

Nem sempre é fácil, contudo, identificar as regras aplicáveis. O Comitê de Supervisão, criado pela Meta para atuar como uma instância independente de controle das decisões de conteúdo no Facebook e no Instagram, por exemplo, por várias vezes criticou o fato de que a regra aplicada a um determinado caso não foi comunicada apropriadamente ao usuário[3]. Muitas vezes, as regras são anunciadas em outras partes da plataforma, como em blogs oficiais, páginas de ajuda, ou até mesmo em declarações à imprensa. Isso pode criar uma verdadeira colcha de retalhos do ponto de vista das regras aplicáveis aos conteúdos de terceiros.

Para lidar com o volume oceânico de conteúdo que precisa ser moderado, plataformas fazem uso de ferramentas de automação. Essas ferramentas são usadas de múltiplas formas. Por exemplo, podem fazer uma triagem do conteúdo segundo a urgência e a gravidade da potencial infração, identificar conteúdo já avaliado como violador das políticas de conteúdo, ou até mesmo tomar decisões com relação a conteúdos e contas.

Não temos informações detalhadas sobre como essas ferramentas são empregadas em grandes plataformas. Sabemos, no entanto, que existem diferentes tipos de ferramentas, com correspondentes limitações. Quando falamos de conteúdo textual, ferramentas de automação têm limitações para distinguir a intenção e o contexto que dão sentido a um post[4]. O Perspective, um produto oferecido pela Jigsaw (do conglomerado Alphabet, que inclui o Google), por exemplo, mostrou dificuldade para discernir quando linguajares ofensivos são apropriados por pessoas LGBTQ+ como forma de responder à opressão[5].

Quando falamos de imagens (estáticas e vídeo) ou áudio, outras limitações se somam. Aqui se torna importante falar de duas grandes categorias

Imagem 4 - Ilustração por Eric Chen.

de ferramentas: as que identificam conteúdo violador já conhecido e as que identificam conteúdo violador ainda não conhecido. Para ferramentas na segunda categoria, que muitas vezes se baseiam em técnicas de *hashing* perceptivo (que compara dois arquivos a partir de impressões digitais para aferir se são suficientemente equivalentes), limitações na aferição de contexto e intenção também se verificam[6]. Ferramentas do primeiro tipo apresentam também problemas próprios, para além do contexto e intenção. Pessoas que busquem contornar sistemas de moderação podem introduzir alterações num determinado conteúdo que evitam sua identificação como conteúdo violador, mesmo quando essas alterações não são perceptíveis para humanos, ou, mesmo quando perceptíveis, ainda permitem a compreensão do conteúdo. Isso ensejaria falsos negativos, evitando que o conteúdo violador fosse removido ou de outra forma restringido.

Outro tipo de desafio é o que se chama de sabotagem do *hash* (*hash poisoning*), por meio do qual um conteúdo não violador é manipulado para que seja identificado pela ferramenta automatizada como se fosse conteúdo violador, resultando em falsos positivos. Esse tipo de ataque pode ser usado para inviabilizar todo o sistema de moderação caso a quantidade de falsos positivos saia do controle. Também pode permitir que pessoas mal-intencionadas prejudiquem determinados grupos de usuários —por exemplo, fazendo com que a imagem de uma candidata à presidência seja removida por ser confundida pela ferramenta com um conteúdo terrorista.

Essas limitações de ferramentas de automação também existirão — e podem ser amplificadas. Além de texto, imagens e áudio, no metaverso também ações de avatares podem violar de políticas de conteúdo, como gestos ofensivos, por exemplo. A violação das políticas de conteúdo também

Imagem 5 - Ilustração por Eric Chen.

pode resultar não diretamente de um objeto digital criado por usuários, mas de sua interação com um dispositivo de AR. Essa criação de objetos e de ambientes é outro fator que multiplica os desafios da moderação. No caso de ambientes de realidade virtual ou realidade aumentada, a imersão pode fazer com que usuários se sintam mais violados (em comparação com como se sentiriam em outro tipo de ambiente), porque percebem de forma diferente ataques aos avatares criados à sua semelhança[7].

Ao contrário de redes sociais como YouTube e TikTok, em que o conteúdo gerado por usuários em geral é permanente e consumido assincronamente, no metaverso muitas vezes o conteúdo será efêmero e consumido sincronamente. Isso limita a janela de oportunidade em que a moderação de conteúdo pode ser realizada; os efeitos lesivos de uma interação podem se dar imediatamente, o que poderia exigir que a moderação fosse realizada de forma contemporânea à própria geração do conteúdo. A natureza efêmera das interações no metaverso também pode limitar a própria identificação de violações, porque em muitas ocasiões pode não ser viável aos usuários envolvidos registrar o ocorrido. Para lidar com essa questão, a Meta propõe manter registros contínuos de pequenos períodos, que usuários poderiam usar para fazer denúncias. Essa saída, que pode não ser suficiente para lidar com todo tipo de denúncia, também não deixa de levantar outras questões, como relacionadas à privacidade dos usuários, particularmente quando ambientes do metaverso reproduzem espaços que as pessoas veem como privados, como salas de aula ou consultórios médicos.

Imagem 6 - Ilustração por Michal Bednarski.

PARTE III

As oportunidades do metaverso

Se a moderação de conteúdo no metaverso apresenta desafios, também traz oportunidades, que têm recebido menos atenção. O metaverso pode aproximar políticas de conteúdo e práticas de moderação de conteúdo das pessoas que são mais diretamente afetadas, mitigando o distanciamento e a centralização que testemunhamos atualmente. Isso certamente não é algo garantido, mas é importante ter essa janela em conta tanto quando buscamos compreender as potencialidades do metaverso quanto quando pensamos em regulação.

Muitas das propostas para o metaverso e até mesmo ambientes já existentes buscam proporcionar espaços diversos. Isso contrasta com a realidade das maiores redes sociais hoje, em que a moderação de conteúdo funciona com a premissa de uma aplicação uniforme de políticas globais[8]. Essa pretensão de submeter todo o mundo a um mesmo conjunto de regras[9] gera uma série de problemas entre os mais discutidos atualmente sobre moderação de conteúdo. Do ponto de vista da aplicação das políticas de conteúdo, significa que as pessoas encarregadas por exercer a moderação ou revisar decisões de moderação muitas vezes não falam a língua do conteúdo objeto da moderação, ou, mesmo quando falam, não têm a compreensão do contexto cultural necessária para estabelecer o propósito e a potencial lesividade de determinado conteúdo.

Do ponto de vista da formulação dessas políticas, a pretensão de submeter todo o mundo a um único conjunto de regras significa um inevitável desalinhamento entre as expectativas de grupos diferentes entre as bilhões de pessoas que têm compreensões também diferentes sobre o que deve ser protegido pela liberdade de expressão. Kate Klonic, nesta toada, narra uma trajetória do desenvolvimento de processos e políticas de conteúdo inicialmente marcado pela cultura jurídica estadunidense quanto à liberdade de expressão, que, por exemplo, muitas vezes contrasta com a cultura de países da União Europeia[10].

No metaverso, no entanto, em vez de ambientes discursivos em que há poucas indicações contextuais para que usuários possam calibrar seu comportamento — quando pensamos num feed de redes como Instagram, Twitter e TikTok —, as pessoas poderão se guiar pelo tipo de espaço que a realidade virtual proporciona. Assim, por exemplo, uma sala de aula sugere certo comportamento, muito diferente daquele sugerido pela arquibancada ou a pista de um *show*. Na sala de aula esperamos nossa vez de falar, não nos levantamos a qualquer momento, não vagamos pelo recinto e de forma

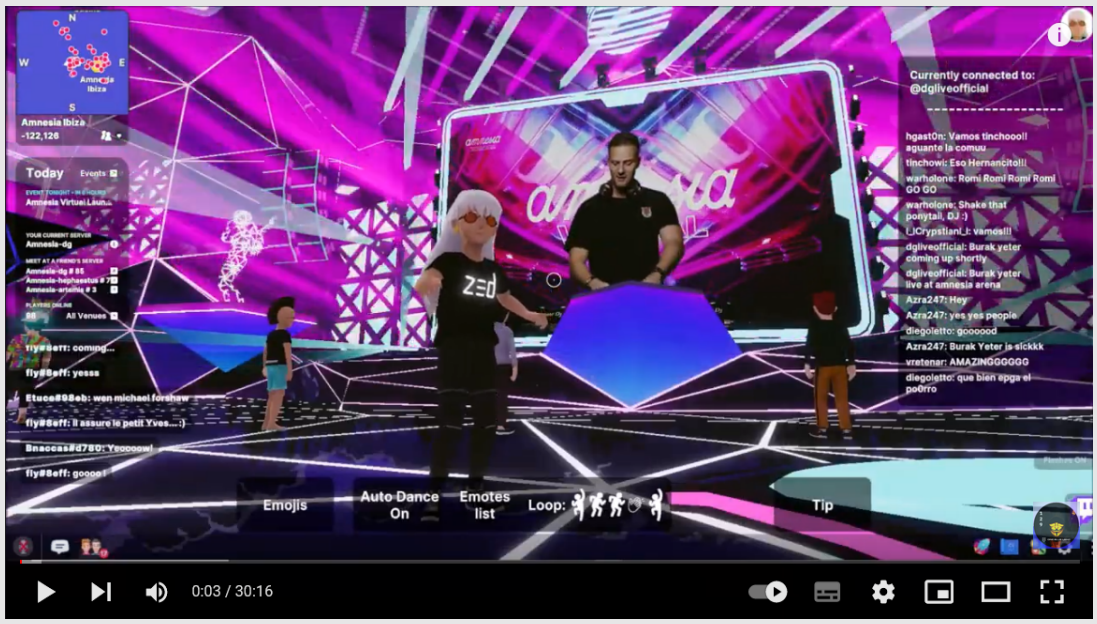


Imagem 7 - Exemplo de show um no metaverso: Evento Mega Festa no Super Show do Metaverso Abertura do primeiro Super Clube Virtual Decentraland.

Imagem 8 - Ilustração por Lin Fritz.

geral adotamos uma conduta atenta ao propósito de aprendizado daquele espaço. Num *show*, a conduta esperada é outra. Podemos mudar de lugar desde que não invadamos o espaço de outras pessoas, podemos conversar desde que não atrapalhemos os demais, etc. Ainda há regras, é claro, mas elas têm outro norte.

O metaverso pode proporcionar espaços desse tipo, que facilitam que as próprias pessoas se governem mesmo sem consultar políticas de conteúdo. Esses espaços também permitem que os usuários saibam o que podem esperar ali — ainda que não completamente, certamente mais do que em comparação com redes sociais como Facebook. Em resumo, os espaços virtuais do metaverso modelados a partir de espaços analógicos a respeito dos quais temos compreensão consolidada também poderão trazer consigo as expectativas estabelecidas para esses espaços.

Ao mesmo tempo, essa pluralidade de espaços também significa que as pessoas têm mais possibilidades de interagir, criar e acessar conteúdo na internet. Para retomar os exemplos, uma sala de aula permite uma dinâmica

de convívio própria; abre espaço para relações especiais de aprendizado; proporciona um fórum para discussões guiadas pelo propósito pedagógico e pela disciplina em questão. Nesse ambiente, as regras podem ser mais rígidas não só do que o estabelecido pela lei, mas também do que é esperado em ambientes não acadêmicos: não basta não violar o direito de ninguém, mesmo a falta de atenção enquanto outra pessoa fala pode ser uma conduta que desrespeita as regras. Em parte, esses ambientes são *constituídos* por essas regras. Se proporcionar a possibilidade de criar e governar esses espaços, o metaverso poderá favorecer a liberdade de expressão permitindo que a moderação de conteúdo seja ajustada ao que as pessoas desejam[11].

Por essa perspectiva, o metaverso também pode trazer a moderação de conteúdo mais para perto das pessoas no Sul Global. Atualmente, ainda que busquem fazer consultas a *stakeholders* de diferentes regiões, as principais plataformas seguem tomando decisões para todo o mundo a partir de suas sedes globais, em geral no Vale do Silício, nos Estados Unidos[12]. Não podemos esperar que o metaverso elimine essa centralização, já que é plausível imaginar que ainda haverá uma expectativa de um piso que não deve ser desrespeitado por nenhuma comunidade, a despeito de suas preferências. Ainda assim, essas comunidades podem ter mais poder e ferramentas de moderação de conteúdo do que têm hoje.

Esse último ponto deve ser ressaltado. O quanto o metaverso pode aproximar as políticas de conteúdo de comunidades no Sul Global dependerá em parte do arranjo resultante da interface das políticas globais da plataforma com as políticas de cada comunidade. Isso é mais evidente quanto à formulação das políticas: uma política global centralizadora, que não deixe espaço para que as comunidades definam suas próprias regras, poderá repetir o que assistimos hoje nas grandes plataformas; caso não

Imagem 9 - Ilustração por Emans.

sejam realizadas consultas regionais, o distanciamento pode ser ainda maior.

Um segundo aspecto é menos evidente: mesmo que as comunidades tenham espaço para formular suas próprias regras, essa aproximação da moderação de conteúdo e o empoderamento de comunidades do Sul Global também dependerá em parte das ferramentas colocadas à disposição dessas comunidades. Comunidades que não sejam capazes de implementar suas próprias políticas poderão na prática ser inviabilizadas como um espaço de governança delegada e localizada.

Por isso, um movimento importante para países como o Brasil pode ser insistir que moderadores comunitários tenham ferramentas que os habilitem a agir com eficiência, em escala e com granularidade — como hoje esperamos que as plataformas sejam capazes de agir. Um exemplo positivo aqui pode ser o AutoModerator à disposição de comunidades no Reddit (os subreddits), que permite que moderadores comunitários sejam capazes de governar subreddits com considerável volume de conteúdo, o que pode lhes proporcionar uma situação superior à de administradores ou moderadores de grupos no Facebook[13], por exemplo[14].

Imagem 10 - Ilustração por Yulong Lli.

Conclusão

As promessas do metaverso são muitas, e ainda é cedo para entender o que é fantasia e o que pode se tornar realidade. Em termos de moderação de conteúdo, o metaverso pode testemunhar uma continuação de debates e dificuldades que já temos hoje, com desafios adicionais relacionados à multiplicidade de tipos de conteúdo e à efemeridade.

Também pode, contudo, ser explorado para favorecer outras configurações na governança das redes sociais, menos centralizadas em decisões corporativas emanadas do Norte Global. A janela de oportunidade para estabelecer essas novas estruturas pode ser breve. Em entrevista sobre o metaverso, um executivo da Meta afirmou que "acertar" na moderação de conteúdo pode ser crucial ao sucesso dos produtos da empresa, porque uma experiência negativa inicial pode ser o suficiente para que as pessoas os rejeitem[15]. Ao mesmo tempo, uma vez que um modelo se estabeleça, será muito mais difícil conseguir alterá-lo, especialmente se as expectativas de usuárias e usuários já tiverem se consolidado.

Referências

1
PARK, Sang-Min ; KIM, Young-Gab, A metaverse: taxonomy, components, applications, and open challenges, **IEEE Access**, v. 10, p. 4209–4251, 2022, p. 4210. **P.8**

2
ONDREJKA, Cory, Escaping the gilded cage: user created content and building the metaverse, **New York Law School Law Review**, v. 49, n. 1, 2004, p. 82–73. **P.8**

3
Ver OVERSIGHT BOARD, Case decision 2020-004-IG-UA (caso do post brasileiro sobre câncer de mama no Instagram) (observando que a aplicação das políticas do Facebook ao Instagram não é claramente comunicada); OVERSIGHT BOARD, Case decision 2020-006-FB-FBR (caso do post em grupo fechado sobre decisão das autoridades de saúde da França a respeito do uso hidroxicloroquina no tratamento de Covid-19) (notando que as políticas de conteúdo estavam espalhadas numa "colcha de retalhos"); OVERSIGHT BOARD, Case decision 2021-001-FB-FBR (caso da suspensão de Donald Trump) (reiterando que informações importantes sobre moderação de conteúdo estavam espalhadas). **P.11**

4
DUARTE, Natasha; LLANSÓ, Emma ; LOUP, Anna, **Mixed messages? The limits of automated social media content analysis,** Washington, DC: Center for Democracy; Technology, 2017; LLANSÓ, Emma et al, **Artificial intelligence, content moderation, and freedom of expression**, Amsterdam: IViR, 2020, p. 3–5. **P.11**

5
OLIVA, Thiago Dias; ANTONIALLI, Dennys Marcelo ; GOMES, Alessandra, Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online, **Sexuality & Culture,** v. 25, n. 2, p. 700–732, 2020. **P.11**

6
GORWA, Robert; BINNS, Reuben ; KATZENBACH, Christian, Algorithmic content moderation: Technical and political challenges in the automation of platform governance, **Big Data & Society**, v. 7, n. 1, p. 2053951719897945, 2020; SHENKMAN, Carey; THAKUR, Dhanaraj ; LLANSÓ, Emma, **Do you see what I see? Capabilities and limits of automated multimedia content analysis**, Washington, DC: Center for Democracy & Technology, 2021.**P.12**

7
CASTRO, Daniel, **Content moderation in multi-user immersive experiences: AR/VR and the future of online speech**, Washington, DC: Information Technology; Innovation Foundation, 2022, p. 15. Disponível em: https://itif.org/publications/2022/02/28/content-moderation-multi-user-immersive-experiences-arvr-and-future-online/. Acesso em 06/10/2022. **P.13**

8
Cf. NITRINI, Rodrigo Vidal, **Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas**, Belo Horizonte: Dialética, 2021, p. 21, sobre o Facebook: "... a política de moderação de conteúdo do Facebook buscava nada menos do que traça regras – de aplicação global, é importante repisar – para a identificação de limites ao 'legítimo' exercício da liberdade de expressão em sua plataforma...". **P.15**

9
Sobre o processo de formulação, ver KETTEMANN, Matthias C ; SCHULZ, Wolfgang, Setting rules for 2.7 billion. A (first) look into Facebook's norm-making system. Disponível em: https://leibniz-hbi.de/en/publications/setting-rules-for-2-7-billion-a-first-look-into-facebook-s-norm-making-system. Acesso em 06/10/2022. **P.15**

10
KLONIC, Kate, The new governors: the people, rules, and processes governing online speech, **Harvard Law Review**, v. 131, n. 6, p. 1598–1670, 2018-04, p. 1621. **P.15**

11
Para uma exposição mais longa desse argumento sobre como a moderação de conteúdo pode ser vista não apenas como limitação da liberdade de expressão, mas como uma manifestação desse direito, ver MONTEIRO, Artur Pericles Lima, Armadilhas à liberdade de expressão na MP 1068/2021, **Jota**, 2021; ver também MONTEIRO, Artur Pericles Lima et al, **Armadilhas e caminhos na regulação da moderação de conteúdo,** São Paulo: InternetLab, 2021, p. 16–17. **P.17**

12
Sobre esses processos no Facebook, novamente ver KETTEMANN ; SCHULZ, Setting rules for 2.7 billion. A (first) look into Facebook's norm-making system. P.17

13
O Facebook passou a oferecer mais ferramentas para grupos. Cf. ARCHEGAS, João Victor; CONCEIÇÃO, Lucas Henrique, **Moderação de conteúdo em grupos brasileiros no Facebook**. Rio de Janeiro: ITS Rio, 2022. Disponível em: <https://itsrio.org/pt/publicacoes/moderacao-de-conteudo-em-grupos-brasileiros-no-facebook/>. P.18

14
Sobre o AutoModerador, ver WRIGHT, Lucas, Automated platform governance through visibility and scale: on the transformational power of Automoderator, **Social Media + Society**, v. 8, n. 1, p. 205630512210770, 2022. P.18

15
ROETTGERS, Andrew Bosworth on Meta's next big challenge: Harassment in the metaverse: *"How much does Meta's plan of getting a billion people to use the metaverse within the next decade depend on getting safety right from the get-go? I think it's hugely important. If the mainstream consumer puts a headset on for the first time and ends up having a really bad experience, that's obviously deleterious to our goals of growing the entire ecosystem. I don't think this is the kind of thing that can wait"*. P.19

Bibliografia

ARCHEGAS, João Victor; CONCEIÇÃO, Lucas Henrique, **Moderação de conteúdo em grupos brasileiros no Facebook.** Rio de Janeiro: ITS Rio, 2022. Disponível em: <https://itsrio.org/pt/publicacoes/moderacao-de-conteudo-em-grupos-brasileiros-no-facebook/>.

CASTRO, Daniel. **Content moderation in multi-user immersive experiences: AR/VR and the future of online speech**. Washington, DC: Information Technology; Innovation Foundation, 2022. Disponível em: <https://itif.org/publications/2022/02/28/content-moderation-multi-user-immersive-experiences-arvr-and-future-online/>.

DUARTE, Natasha; LLANSÓ, Emma ; LOUP, Anna. **Mixed messages? The limits of automated social media content analysis.** Washington, DC: Center for Democracy; Technology, 2017.

GORWA, Robert; BINNS, Reuben ; KATZENBACH, Christian. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. **Big Data & Society,** v. 7, n. 1, p. 2053951719897945, 2020.

KETTEMANN, Matthias C ; SCHULZ, Wolfgang. Setting rules for 2.7 billion. A (first) look into Facebook's norm-making system.

KLONIC, Kate. The new governors: the people, rules, and processes governing online speech. **Harvard Law Review**, v. 131, n. 6, p. 1598–1670, 2018-04.

LLANSÓ, Emma; HOBOKEN, Joris van; LEERSSEN, Paddy; et al. **Artificial intelligence, content moderation, and freedom of expression**. Amsterdam: IViR, 2020.

MONTEIRO, Artur Pericles Lima. Armadilhas à liberdade de expressão na MP 1068/2021. **Jota,** 2021.

MONTEIRO, Artur Pericles Lima; CRUZ, Francisco Brito; SILVEIRA, Juliana Fonteles da; et al. **Armadilhas e caminhos na regulação da moderação de conteúdo.** São Paulo: InternetLab, 2021. Disponível em: <https://www.internetlab.org.br/pt/liberdade-de-expressao/armadilhas-caminhos-moderacao/>.

NITRINI, Rodrigo Vidal. **Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas.** Belo Horizonte: Dialética, 2021.

OLIVA, Thiago Dias; ANTONIALLI, Dennys Marcelo ; GOMES, Alessandra. Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online. **Sexuality & Culture**, v. 25, n. 2, p. 700–732, 2020.

ONDREJKA, Cory. Escaping the gilded cage: user created content and building the metaverse. **New York Law School Law Review**, v. 49, n. 1, 2004. (81-102).

OVERSIGHT BOARD. Case decision 2020-004-IG-UA. Disponível em: <https://oversightboard.com/decision/IG-7THR3SI1/>. Acesso em: 22 jun. 2022.

OVERSIGHT BOARD. Case decision 2020-006-FB-FBR. Disponível em: <https://www.oversightboard.com/decision/FB-XWJQBU9A/>. Acesso em: 21 jun. 2022.

OVERSIGHT BOARD. Case decision 2021-001-FB-FBR. Disponível em: <https://oversightboard.com/decision/FB-691QAMHJ/>. Acesso em: 22 jun. 2022.

PARK, Sang-Min ; KIM, Young-Gab. A metaverse: taxonomy, components, applications, and open challenges. **IEEE Access**, v. 10, p. 4209–4251, 2022.

ROETTGERS, Janko. Andrew Bosworth on Meta's next big challenge: Harassment in the metaverse. **Protocol,** 2021. Disponível em: <https://www.protocol.com/vr-harassment-metaverse-bosworth>.

SHENKMAN, Carey; THAKUR, Dhanaraj ; LLANSO, Emma. **Do you see what I see? Capabilities and limits of automated multimedia content analysis.** Washington, DC: Center for Democracy & Technology, 2021.

WRIGHT, Lucas. Automated platform governance through visibility and scale: on the transformational power of Automoderator. **Social Media + Society,** v. 8, n. 1, p. 205630512210770, 2022.

SOBRE O AUTOR

Artur Pericles Lima Monteiro

Doutor em direito pela Universidade de São Paulo (USP), é Associate Research Scholar na Yale Law School (YLS) e Wikimedia Fellow no Information Society Project/YLS. Também é integrante do grupo constituição, instituições & política, da USP.

Acesse nossas redes

itsrio.org