A Safe Space for Everyone – A Plea for a Democratic and Participative Metaverse

Octavia Madeira & Georg Plattner

Malevolent Actors in the Metaverse

The vision of a metaverse presented by Meta CEO Mark Zuckerberg in October 2021 was a watershed moment for society and for the tech world. Although the concept of a second digital life, including a digital identity, is neither new nor exclusive to Meta (see, for example, Second Life), this was the first time that almost all the possible functionalities of the metaverse had been presented up to that point. Here, there is a special emphasis on immersion via virtual reality which, as an extension of today's Internet applications, is meant to give users a completely new sense of participation and let them experience the metaverse in a multimodal and multi-sensory way. The presentation of the vision also focused on technological permeability, on the diffusion of social media in all areas of human life and therefore on the displacement of social media as a purely entertainment platform.

Should the metaverse turn out to be as Zuckerberg and other proponents envision it, this would mean a radical transformation of social interaction with the digital space, and also a radical change in our everyday lives. Shopping could increasingly shift to the metaverse as an immersive experience, sports classes could take place in a virtual environment and virtual church services could be held with believers from all over the world. The world of work has already permanently changed, partly due to the coronavirus pandemic – and we could soon move from working from home to working in the meta-office.

But these innovations will not only change our everyday lives – they will also cause extremism and radicalisation to strike out in new directions and transform to adapt to new environments. Extremists use technologies that are cheap, readily available, easy to use and widely accessible for their purposes, like propaganda, communication and recruitment. Using technology for a function other than that intended by the developers with the intention of doing harm to others is an inherently creative process.

Cropley, Kaufman and Cropley (2008) call this "malevolent creativity". They define it as a form of creativity that "is deemed necessary by some society, group, or individual to fulfil goals they regard as desirable, but has serious negative consequences for some other group, these negative consequences being fully intended by the first group" (p. 106). We describe actors who display malevolent creativity (such as extremists or spreaders of fake news) as malevolent actors.

In the past, malevolent actors were very creative especially when it came to realigning their own organisation and distributing their own ideology. The digital revolution has equipped them with an unprecedented number of tools with which to further their cause: from (encrypted and instant) mass communication for propaganda and recruitment to alternative instruments for financing operations and logistics through to new means of destruction and terror. Recent technological advances have opened up a wide range of new opportunities for malevolent actors. For example, Web 2.0, the rise of social media and the availability of nearly all content on the Internet have enabled these actors to easily connect with other like-minded individuals and form almost entirely closed communities that reinforce their own views.

Research into the metaverse as the successor to social media and the mobile Internet can provide important insights into how malevolent actors could creatively use the metaverse. While we generally agree with Joe Whittaker and others (Whittaker, 2022; Valentini, Lorusso & Stephan, 2020) that distinguishing between offline and online radicalisation does not make sense from an analytical perspective, the way in which malevolent actors currently use social media could give an idea of the metaverse of the future.

It is generally recognised that malevolent actors (with different ideological backgrounds) began to make use of the Internet and its possibilities at an early stage (Feldman, 2020; Fisher, 2015; Stewart, 2021; Lehmann & Schröder, 2021). They used new technologies in creative ways in order to evade monitoring and detection and also to improve their own operations. As an anonymous place of countless possibilities where one can find a wealth of information tailored to one's own interests, the Internet is a gold mine for extremists (Bertram, 2016, p. 232).

While research on the radicalisation patterns of convicted jihadi terrorists has shown that offline networks played a much greater role in their radicalisation than online networks (Hamid & Ariza, 2022), other research indicates that the Internet has a more important role for right-wing extremists. This applies especially to the planning of their attacks and actions (von

Behr et al., 2013; Gill et al., 2017). "The Internet is largely a facilitative tool that affords greater opportunities for violent radicalization and attack planning. Nevertheless, radicalization and attack planning are not dependent on the Internet […]." (Gill et al. 2017, p. 113).

Social media has been used by malevolent groups to create, target and distribute self- generated content without the traditional processes of vetting used by traditional media companies and while avoiding policing and censorship from nation states (Droogan et al., 2018, p. 171). Furthermore, social media has also become an instrument of social interaction for those who are already radicalised and those whom they want to convince or who are interested in their activities (Conway, 2017).

The introduction of the metaverse could further reinforce this momentum. By further bridging the gap between offline and online, it could be even more difficult to maintain the distinction between the two spheres of radicalisation (and extremism and terrorism). At present, offline networks provide familiarity and a close environment and are more likely to evade security services than online extremists (Hamid & Ariza, 2022). The future metaverse could bring together these advantages of the offline world in an extensive and immersive digital experience. Combined with the advantages of the online world – instant mass communication and propaganda – the metaverse could become an even bigger game-changer than the Internet and social media were.

The Metaverse as a Democratic Space

The metaverse is still in the early stages of development and has a long way to go before it reaches a certain stage of maturity in which promises, and actual functionalities are implemented. It is already apparent that the risks of the metaverse are comparable to those of social media and, in the past, a response often came too late. Freedom and security will probably be the decisive variables in this technology of the future, which makes engaging with malevolent actors all the more crucial (Neuberger, 2023). In the initial phase of the metaverse, it is already becoming clear that malevolent actors are finding fertile ground – as illustrated, for example, by incidents of sexual harassment that have occurred in the current test versions of the metaverse (Bazu, 2021; Bovermann, 2022; Diaz, 2022; Wiederhold, 2022).

How can these developments be tackled? How can they be prevented before they cause harm? It will be important to ensure the democratic

involvement of actors and marginalised groups in decision-making and development processes. While this would now be a genuinely reactive process in the case of social media, the developers of the metaverse still could build beneficial structures. Democratisation of social media is desirable from a sociopolitical perspective because it is a powerful tool due to its widespread use and its economic and cultural importance. This power should be democratically legitimised and controlled (Engelmann et al., 2020). However, democratic safeguarding should not follow a party-political pattern.

In the development of the metaverse, social media should be informative in various ways – from the creativity with which malevolent actors use new media and technologies (see above) through to the democratic involvement of users. Social media operators have already tried to take account of the aspect of participation:

- META conceived the idea of an Oversight Board[1] in 2018 as a body whose independent judgement could help the company make tough content decisions. This board is committed to being independent, accessible, and transparent. META has granted it the authority to decide whether content should be allowed or removed.
- Twitter has been advised by a Trust and Safety Council in the past. This consisted of various NGOs and researchers who advised the company on online security issues. Elon Musk dissolved the Council after taking over the company (The Associated Press, 2022).
- On its YouTube video platform, Google has introduced the Priority Flagger Programme[2]. This enables NGOs and public authorities to use highly effective tools to report content that violates the Community Guidelines. This flagged content is then reviewed by moderators as a priority. However, the deletion criteria are the same as for any other reports. The programme was revised by YouTube in 2021, which led to major criticism from the community (Meineck, 2021).

In general, there seems to be a worrying trend on social media to cut back on these participative models of moderation and security in favour of artificial intelligence (AI) applications (Gorwa et al., 2020; Llansó, 2020). However, AI solutions cannot and should not replace the involvement of civil society in decision-making processes and questions of democratic culture, not least because AI-supported content moderation solutions are

1 https://www.oversightboard.com/.
2 https://support.google.com/youtube/answer/7554338?hl=de.

still prone to error and lack transparency (Gillespie, 2020; Gorwa et al., 2020).

Encouraging Participation and Democratic Involvement

In social media research and particularly in platform governance research, important approaches can be found that may help to enable a democratic and inclusive metaverse. In addition to essential cooperation between operators and governmental and non-governmental actors on issues of transparency and research, there is an emphasis on actively strengthening democratic actors and narratives (Bundtzen & Schwieter, 2023; Engelmann et al., 2020; Rau et al., 2022).

This strategy is crucial to ensure that a state's repressive apparatus is actually only used as a measure of last resort to stop malevolent actors. Democratic argument and discourse must be possible in an inclusive metaverse without people constantly having to fear repression and restriction. Instead, platform operators can also take steps in the metaverse to consciously and actively promote democratic actors and narratives and thus build democratic resilience in the metaverse.

Here too, the metaverse can take inspiration from existing approaches in the social media field, such as YouTube's trusted flagging programme. Democratic actors, e.g. NGOs and government organisations, specialising in areas such as hate speech, group-focused enmity or strengthening democracy could have access to special reporting tools. They could also be given extended powers to contextualise questionable content.

However, as well as reinforcing democratic narratives, the democratisation of the platform itself is a crucial factor for inclusivity. Involving users in decision-making and design processes can have enormous added value for a platform that is interested in democratic interaction. Marginalised groups and their representatives know exactly where hate and harassment may be lurking in the digital space. By involving such stakeholders at an early stage, some of the mistakes that were made on social media could be minimised from the outset.

In political practice, mini publics have already proved effective as an instrument of user participation (Escobar & Elstub, 2017; Smith & Setälä, 2018). Mini publics are groups of (randomly or systematically) selected citizens who work together over an extended period to examine socially relevant issues, with the inclusion of external sources, e.g., scientific exper-

tise. Topics are examined, discussed and assessed from a broad range of perspectives, and the resulting recommendations are forwarded to political decision-makers (Escobar & Elstub, 2017; Pek et al., 2023). One example of this is the virtual citizens' assembly in Germany. In June 2022, its members debated the consequences of using artificial intelligence (Buergerrat.de, 2022). These types of assemblies allow platform-specific topics to be discussed with the aim of ensuring that decision making is more democratic.

Although quite controversial (see above), platform councils can also develop potential for promoting democracy if they are able to operate independently, objectively and transparently (Haggart & Keller, 2021; Rau et al., 2022). To ensure this, platform councils of this type could be based on the press and broadcasting councils that are already established in Germany, in line with the recommendations of Kettemann and Fertmann (2021). It should be noted, however, that responsibilities (geographical, practical), participants (citizens, experts, NGOs, political decision-makers), and not least powers (quasi-judicial, advisory) must be part of the social discourse and cannot yet be conclusively clarified (Cowls et al., 2022; Kettemann & Fertmann, 2021). Furthermore, such councils could boost public confidence – the more diverse and transparent their line-up is and the more publicly visible the effects of their recommendations are.

Finally, the aim must also be to strengthen media literacy and policy competence by means of various training opportunities. These should be designed in such a way that individuals who are not (or no longer) associated with the education system are also able to benefit from them. Here it is vital to provide the necessary tools for dealing with fake news, other manipulated or extremist content and hate speech on the Internet. One example to mention is the Good Gaming – Well Played Democracy project[3] directed by the Amadeu Antonio Foundation, which aims to raise the gaming community's awareness of extremist content, among other things.

In addition, it must be noted that building a democratic metaverse is not solely a task for citizens. The creation of a digital twin in the sense of a well-fortified democracy is also important. However, according to Rau et al. (2022), this does not exclusively mean the use of repressive measures such as deletion or suppression of problematic content (see, for example, Bellanova & De Goede, 2022) but also, coupled with this, the strengthening of democratic actors, e.g. through algorithmically increased visibility. In

3 https://www.amadeu-antonio-stiftung.de/projekte/good-gaming-well-played-democracy/.

this context, the empowerment of marginalised democratic actor groups becomes especially important to adequately represent social diversity. They are properly trained to recognise problematic content at an early stage, for example, and can thus also be consulted for advice (Rau et al., 2022). The use of counter speech could also be another strategy for tackling extremist content in the metaverse (Clever et al., 2022; Hangartner et al., 2021; Kunst et al., 2021; Morten et al., 2020). The term (digital) counter speech refers to comments or other content posted as a response to hate speech in order to minimise and weaken the impact of it or to support potential victims (Ernst et al., 2022; Garland et al., 2022). In this regard, studies have shown that counter speech can be an effective means of tackling extremist content and reducing it effectively (Garland et al., 2022; Hangartner et al., 2021). In the context of newer technological complexes, e.g. AI, consideration is currently being given to implementing counter speech automatically in certain circumstances, although final concepts and responsibilities are still the subject of intensive discussion (Clever et al., 2022).

In addition to participatory methods, legislation can also be used to prevent extremist content. In Germany, the dissemination of unconstitutional symbols and signs is forbidden, and perpetrators can be prosecuted. Germany's Network Enforcement Act (Netzwerkdurchsetzungsgesetz, NetzDG) also provides a legal framework for dealing with hate crime on social media. Accordingly, the Terrorist Content Online Regulation (European Union, 2021) requires platform operators offering services in the EU to remove or block reported terrorist content within one hour. Recent results of extremism research indicate, however, that so-called legal but harmful content is already proving to be a major challenge and is likely to be of significance in the metaverse as well (Jiang et al., 2021; Rau et al., 2022). This includes, for example, digital content that may have a subtle radicalising effect but is not unlawful. However, it should be noted in this regard that content moderation must comply with the constitutional principle of free speech. Consequently, it is to be assumed that the ongoing discussion on the relationship between freedom and security will also significantly influence the design of the metaverse and will, or must, be the result of a negotiation process involving society as a whole in order to guarantee the democratic dimension.

Discussion

If the immersiveness of the metaverse lives up to Mark Zuckerberg's vision, it is very likely to have a huge impact on our everyday lives and on social interaction. This immersiveness would mean that the operators of the metaverse (or metaverses) would need to deal intensively with questions of democratisation. Not only would the state probably play a (yet to be defined) role in a metaverse, but its users must also be enabled to participate democratically in it. This would help to make the platform inclusive and as safe as possible from malevolent actors.

Building on the social media research of recent decades, there are many common points of reference which can support and steer the design of a democratic metaverse. As mentioned above, the metaverse is still at an early stage of development. However, given the rapid pace of advancement, it is vital to support this process, stay on the ball and take an active role in discussions. A multi-perspective approach from all stakeholders involved is also relevant to ensuring a balance between security and freedom for all users. The possibilities outlined here for building a metaverse present some solutions for implementing democratic pillars. In summary, the following solutions should be deployed by platform operators:

- early implementation of methods for user participation, e.g. mini-publics or independent platform councils
- strengthening of democratic actors and inclusion of marginalised groups
- reference to existing scientific research findings on social media, hate speech and (digital) extremism, as well as open cooperation with research institutions
- offer of educational opportunities in cooperation with democratic actors

Final and concrete implementation is still currently the subject of lively discussion. However, the status of the early development phase of the metaverse is encouraging active participation, which is also reflected in this Immersive Democracy Project and can be understood as an invitation to this process. Participation is not a panacea for the dangers lurking in the digital space. But it is an important source of support that can help to empower marginalised groups or individuals in specific ways and thus give them the tools to work together with operators against discrimination and hate in the metaverse. Now is the time to develop these tools and to make sure that a future metaverse is as safe and secure as possible for everyone.

References

Bazu, T. (2021). The metaverse has a groping problem already. *MIT Technology Review*. https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/

von Behr, I., Reding, A., Edwards, C., & Gribbon, L. (2013). *Radicalisation in the digital era: The use of the internet in 15 cases of terrorism and extremism*. RAND Europe.

Bellanova, R., & De Goede, M. (2022). Co-producing security: Platform content moderation and European security integration. *JCMS: Journal of Common Market Studies*, *60*(5), 1316–1334. https://doi.org/10.1111/jcms.13306

Bertram, L. (2016). Terrorism, the internet and the social media advantage: Exploring how terrorist organizations exploit aspects of the internet, social media and how these same platforms could be used to counter-violent extremism. *Journal for Deradicalization*, *2016*(7), 225–252.

Bovermann, P. (2022). Online-Belästigungen im Metaverse – Am eigenen Leib. *Süddeutsche Zeitung*. https://www.sueddeutsche.de/kultur/metaverse-vr-virtual-reality-microsoft-sexuelle-belaestigung-1.5519527?print=true

Buergerrat.de. (2022). Bürgerrat diskutierte über künstliche Intelligenz. *Buergerrat.de*. https://www.buergerrat.de/aktuelles/buergerrat-diskutierte-ueber-kuenstliche-intelligenz/

Bundtzen, S., & Schwieter, C. (2023). *Datenzugang zu Social-Media-Plattformen für die Forschung: Lehren aus bisherigen Maßnahmen und Empfehlungen zur Stärkung von Initiativen inner- und außerhalb der EU*. Institute for Strategic Dialogue (ISD).

Clever, L., Klapproth, J., & Frischlich, L. (2022). Automatisierte (Gegen-)Rede? Social Bots als digitales Sprachrohr ihrer Nutzer*innen. In J. Ernst, M. Trompeta, & H.-J. Roth (Eds.), *Gegenrede digital: Neue und alte Herausforderungen interkultureller Bildungsarbeit in Zeiten der Digitalisierung* (pp. 11–26). Springer Fachmedien. https://doi.org/10.1007/978-3-658-36540-0_2

Conway, M. (2017). Determining the role of the internet in violent extremism and terrorism: Six suggestions for progressing research. *Studies in Conflict & Terrorism*, *40*(1), 77–98. https://doi.org/10.1080/1057610X.2016.1157408

Cowls, J., Darius, P., Santistevan, D., & Schramm, M. (2022). Constitutional metaphors: Facebook's "supreme court" and the legitimation of platform governance. *New Media & Society*. https://doi.org/10.1177/14614448221085559

Cropley, D. H., Kaufman, J. C., & Cropley, A. J. (2008). Malevolent creativity: A functional model of creativity in terrorism and crime. *Creativity Research Journal*, *20*(2), 105–115. https://doi.org/10.1080/10400410802059424

Diaz, A. (2022). Disturbing reports of sexual assaults in the metaverse: 'It's a free show'. *New York Post*. https://nypost.com/2022/05/27/women-are-being-sexually-assaulted-in-the-metaverse/

Droogan, J., Waldek, L., & Blackhall, R. (2018). Innovation and terror: An analysis of the use of social media by terror-related groups in the Asia Pacific. *Journal of Policing, Intelligence and Counter Terrorism*, *13*(2), 170–184. https://doi.org/10.1080/18335330.2018.1476773

Engelmann, S., Grossklags, J., & Herzog, L. (2020). Should users participate in governing social media? Philosophical and technical considerations of democratic social media. *First Monday*, *25*(12). https://doi.org/10.5210/fm.v25i12.10525

Ernst, J., Trompeta, M., & Roth, H.-J. (2022). Gegenrede digital – Einleitung in den Band. In J. Ernst, M. Trompeta, & H.-J. Roth (Eds.), *Gegenrede digital* (pp. 1–7). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-36540-0_1

Escobar, O., & Elstub, S. (2017). Forms of mini-publics: An introduction to deliberative innovations in democratic practice [Research and Development Note]. *New Democracy*.

European Union. (2021). Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online (Text with EEA relevance). *Official Journal of the European Union*.

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 11(1), Article 3. https://doi.org/10.1140/epjds/s13688-021-00314-6

Gill, P., Corner, E., Conway, M., Thornton, A., Bloom, M., & Horgan, J. (2017). Terrorist use of the internet by the numbers: Quantifying behaviors, patterns, and processes. *Criminology & Public Policy*, 16(1), 99–117. https://doi.org/10.1111/1745-9133.12249

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2). https://doi.org/10.1177/2053951720943234

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Haggart, B., & Keller, C. I. (2021). Democratic legitimacy in global platform governance. *Telecommunications Policy*, *45*(6). https://doi.org/10.1016/j.telpol.2021.102152

Hamid, N., & Ariza, C. (2022). *Offline versus online radicalisation: Which is the bigger threat?* London: Global Network on Extremism & Technology.

Hangartner, D., Schmid, L., Helbling, M., & Zingg, A. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50). https://doi.org/10.1073/pnas.2116310118

Jiang, J. A., Scheuerman, M. K., Fiesler, C., Brubaker, J. R., & Bovet, A. (2021). Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, 16(8). https://doi.org/10.1371/journal.pone.0256762

Kettemann, M. C., & Fertmann, M. (2021). *Die Demokratie Plattformfest Machen: Social Media Councils als Werkzeug zur gesellschaftlichen Rückbindung der privaten Ordnungen digitaler Plattformen*. Friedrich-Naumann-Stiftung.

Kunst, M., Porten-Cheé, P., Emmer, M., & Eilders, C. (2021). Do "Good Citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology & Politics*, 18(3), 258–273. https://doi.org/10.1080/19331681.2020.1871149

Llansó, E. J. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720920686

Meineck, S. (2021). Trusted Flagger: YouTube serviert freiwillige Helfer:innen ab. *netzpolitik.org*. https://netzpolitik.org/2021/trusted-flagger-youtube-serviert-freiwillige-helferinnen-ab/

Morten, A., Frischlich, L., & Rieger, D. (2020). Gegenbotschaften als Baustein der Extremismusprävention. In J. B. Schmitt, J. Ernst, D. Rieger, & H.-J. Roth (Eds.), *Propaganda und Prävention* (pp. 581–589). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-28538-8_32

Neuberger, C. (2023). Sicherheit und Freiheit in der digitalen Öffentlichkeit. In N. J. Saam & H. Bielefeldt (Eds.), *Sozialtheorie* (pp. 297–308). transcript Verlag.

Pek, S., Mena, S., & Lyons, B. (2023). The role of deliberative mini-publics in improving the deliberative capacity of multi-stakeholder initiatives. *Business Ethics Quarterly*, 33(1), 102–145. https://doi.org/10.1017/beq.2022.20

Rau, J., Kero, S., Hofmann, V., Dinar, C., & Heldt, A. P. (2022). Rechtsextreme Online-Kommunikation in Krisenzeiten: Herausforderungen und Interventionsmöglichkeiten aus Sicht der Rechtsextremismus- und Platform-Governance-Forschung. *Arbeitspapiere des Hans-Bredow-Instituts*. https://doi.org/10.21241/SSOAR.78072

Smith, G., & Setälä, M. (2018). Mini-publics and deliberative democracy. In A. Bächtiger, J. S. Dryzek, J. Mansbridge, & M. Warren (Eds.), *The Oxford Handbook of Deliberative Democracy* (pp. 299–314). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198747369.013.27

The Associated Press. (2022, December 12). Musk's Twitter has dissolved its Trust and Safety Council. *National Public Radio (NPR)*.

Wiederhold, B. K. (2022). Sexual harassment in the metaverse. *Cyberpsychology, Behavior, and Social Networking*, 25(8), 479–480. https://doi.org/10.1089/cyber.2022.29253. editorial