Addressing Hateful and Misleading Content in the Metaverse

Inga K. Trauthig and Samuel C. Woolley

1 The metaverse in 2023

In 2023 the tech sector has been defined by a spate of announcements about job cuts and reverberating concerns about artificial intelligence tools, particularly the chatbot ChatGPT (Alfonseca and Zahn 2023; Shankland 2023). The metaverse—the extended reality (XR) space tied to a decades-old science fiction vision of an immersive virtual reality (VR) world (Kirkpatrick 2022)—is partially linked to both of Silicon Valley's woes. For example, it has consumed substantial resources from social media companies such as Meta, which invested billions into its metaverse business Reality Labs and hence exerted economic pressure on the company to save money elsewhere (Daniel 2022).

Online safety issues in the metaverse may worsen if actions are not taken. We outline our argument and present a call for action over the next sections, which start with explaining current conceptualizations of the metaverse and giving a brief interdisciplinary overview of research on the metaverse. We then deduct existing and potential future avenues for malevolent exploitation with a focus on mis-, and disinformation, and finally outline recommendations that could alleviate negative effects on the individual and societies from the metaverse.

The metaverse presents novel avenues for abuse regarding behavioral change related to embodied experiences, which researchers such as Ahn et al. have been studying for years (Ahn, Bailenson, and Park 2014). We fear that malevolent exploitation of the metaverse is not being adequately addressed by either companies or policymakers. Pundits might point fun at both Meta's XR world and broader tech sector moves toward VR and augmented reality (AR). But the very fact that the most powerful companies in tech have invested billions of dollars into these tools suggests that some business executives hope the investments will pay out and the metaverse will become more popular among consumers (Chohan 2022).

Experts such as Brittan Heller and Katherine Lo continue to underscore the particular dangers of a disembodied digital world (Atlantic Council 2022). Imagine entering a virtual environment that you navigate as an avatar. Not only do you see and hear as the avatar, you feel as it does, too. You are in a meeting space with other users, chatting casually about some topic or another. Suddenly, your avatar is taken by the hand, pulled toward another avatar and touched without your consent. In the offline world, you experience the touch. Offline, you may physically be able to remove yourself from the situation, report to nearby authorities of different kinds, or even find mental health support if the incident is difficult to process. But in the metaverse it is unclear if this incident counts as an assault—and if so, who decides that? Will there be any consequences for the perpetrator? Do you even know who the perpetrator "really"

doi:10.54501/jots.v1i5.109

is? This can be difficult for new users. For example, the brief health and safety videos provided in Meta Quest's Safety Center for all Oculus and Meta Quest models, that we could find, do not address instances of assault and what to do about them (Meta Quest 2022). However, the written guidance based on Meta's Code of Conduct for Virtual Experiences states that "any form of non-consensual intimate activity" is not allowed as it counts as either illegal or abusive behavior; nonetheless, how potentially negative experiences like these can be avoided and addressed is vague given that "persistent or egregious issues not resolved by admins or developers may be directly addressed by Meta as a platform, through actions like limiting developer access or app removal" (Meta Quest 2022). As people, including children, interact in the metaverse, these questions become pressing. But practical resolutions are not always available (Wiederhold 2022), while people are already experiencing harassment over VR (Basu 2021).

1.1 Current understandings of the metaverse

Four main characteristics define the metaverse. First, it has an immersive nature (Bailenson 2018). The metaverse is considered a virtual world that merges multiple different–virtual–dimensions. In the words of Dionisio et al. (2013), it is "an integrated network of 3D virtual worlds." Like the internet, it is a world removed from our physical world on Earth. Yet it is immersive because we immerse a version of ourselves as avatars into those environments, usually through augmented reality (AR) or virtual reality (VR), which coalesce to form extended reality (XR) (McEwan 2021). On a more abstract level, it is also immersive because it strives to integrate education, work, and social contexts into its network structure, and concurrently develops its own digital economy (Chen et al. 2022). Optimistically, these possibilities are promising, as they allow users to access virtual worlds that they might never have been able to access offline. VR has already been widely used to entertain, improve efficacy by providing trainings, support research, and improve healthcare (Wiederhold 2022).

However, and second of the definitional characteristics of the metaverse, it is not utopian, and thus not aiming for an idealistic state of all involved. Instead, it is defined by economic competition of various entities, which include tech giants like Apple, Meta, and Microsoft (Mac, Frenkel, and Roose 2022). But the companies' success depends on their ability to bring virtual and augmented reality tools to far more people. And capitalist competition might actually be an inhibitor to further success—"a *Ready Player One*-like single unified place called 'the metaverse' is still largely impossible" (Ravenscraft 2022), as companies may not perceive interoperability as the most profitable approach.

Third, the metaverse is linked to a notion of pervasiveness—in other words, the removal of existing online-offline binaries. As Heller, a fellow at the Digital Forensic Research Lab focusing on VR, explains, "It will be constantly on. You will not be able to turn it off" (Atlantic Council 2022). Functionally, this dynamic is supported by the fact that different interfaces can connect us to the metaverse: "Virtual worlds—such as aspects of Fortnite that can be accessed through PCs, game consoles, and even phones—have started referring to themselves as 'the metaverse'" (Ravenscraft 2022).

But this leads us to the last, and arguably most relevant, characteristic of the metaverse as it exists in 2023: It's unclear exactly what "metaverse" means. In other words, it can mean different things to different people, with minimal consensus based on the three previous characteristics. Video game players, in particular, are likely to believe that the metaverse has existed for some time already. Gamers have been buying and selling goods in the virtual realms of *World of Warcraft* and socializing in *Second Life* for years now. Currently, Meta is working hard to propagate its vision of the metaverse as an all-encompassing system operating on Horizon Home, whose rollout started in 2022 and which is the new landing point for anyone with a Quest headset. Horizon Home was designed to make it easier for users to just hang out. It is supposed to provide a user-friendly upgrade from interacting via Meta-linked VR apps, such as Horizon Worlds, VR Chat, or Horizon Venues (since June 2022 integrated into Horizon Worlds) (Meta 2021; Silberling 2021). Some, like computer scientist and current CEO of Unanimous AI Louis Rosenberg, echo Heller and underline the centrality of augmented versus virtual reality:

> The true Metaverse—the one that becomes the central platform of our lives—will be an augmented world. If we do it right, it will be magical, and it will be everywhere. (Rosenberg 2022)

However, and critically, it may be the case that any VR or AR developments we currently attach to the metaverse simply define the next wave of internet developments—effectively conveying to us an XR future that combines virtual realities with the real world. Cool VR games and digital avatars in Zoom might be spreading rapidly, but it is possible they will remain something that most of us will still think of as “the internet.” Existing trust and safety issues will therefore carry over as well as take on new shapes (Slater et al. 2019).

1.2 Existing research on the metaverse

Existing research on the metaverse largely focuses on its feasibility, specific features, privacy concerns, and related regulatory implications. Fewer articles analyze the current, and potential future, individual harmful implications for people’s lives and/or structural effects for our societies and political systems. The following sections discuss research published after Mark Zuckerberg’s announcement of Facebook’s rebranding to Meta in October 2021—because this was a pivotal moment wherein the metaverse was supposed to move from the margins to the mainstream.

Existing academic research examines the metaverse’s slow rise in popularity but simultaneously comments on privacy concerns and other regulatory implications of the metaverse. For example, Ma (2022) elaborates on the metaverse’s “transformative opportunities” but simultaneous inherent challenges for mainstream adoption. McEwan (2021) outlines her concerns about Meta’s attempts at dominating the metaverse commercially in light of Meta’s potential abuse of user data, while Berryman and Leaver (2022) expand the prolific field of social media influencer research to the metaverse and argue that there is an urgent need for ethical guidelines. Additionally, Smaili and Rancourt-Raymond (2022) explain how business crime already happens in the metaverse and what future frauds might look like, Casswell (2022) comments on the concerns of alcohol marketing in the metaverse, and Heller and Bar-Zeev (2021) critically discuss the future of advertising.

A comprehensive article by 41 academics with eclectic backgrounds assessed the “Metaverse beyond the hype” and identified the following as among the most pressing relevant research question: “How does misinformation, disinformation, and fake news share and spread in the metaverse?” (Dwivedi et al. 2022). But the dangers posed by the metaverse are not simply confined to content as challenging to parse and moderate as, say, mis- and disinformation. Existing projects focusing on trust and safety issues concentrate on children and/or sexual exploitation, for example. Researchers for The Centre for Countering Digital Hate, for instance, posed as minors and spent time on VR Chat, a virtual world accessible via VR headsets such as Meta Quest but also Windows Mixed Reality or a regular desktop setup (which largely lacks the VR experience), finding that users were “exposed to abusive behavior every seven minutes”; this included bullying or confronting users with sexual content (Louise 2022). In June 2022, Pew asked experts about positive and negative aspects of the embodied digital world. One main finding was that a majority of people are convinced of the metaverse's capacity to accelerate: "These new worlds could dramatically magnify every human trait and tendency—both the bad and the good" (Atske 2022).

Given our team's own research on malevolent exploitation of emerging technologies over the last few years, which we incorporate into the next section, we outline three trends that seem central in shaping existing and future malevolent exploitation of the metaverse: (1) potential commercial success of VR/XR platforms and companies selling related gadgets outside the US; (2) potentially more harmful impacts on the individual due to the metaverse's immersive nature; and (3) increased difficulty in protecting and managing the exposure to hateful and misleading content and behavior in the metaverse (versus established moderation systems on Facebook, for example).

2 What bad actors exploit and how the metaverse fits in

Earlier we discussed Zuckerberg's announcement of reimagining, and with that rebranding, Facebook to Meta as a potential pivotal turning point for moving the metaverse from the margins to the mainstream. If this expansion into the mainstream proves successful and more users spend time in VR/XR realities, the metaverse might be exploited further. With increased usage, emerging technologies also enter the radar of malevolent actors (Doctor, Elson, and Hunter 2022) and those aiming to manipulate public opinion, who have already been noting that the metaverse is an addition to their arsenal (Woolley 2020).

2.1 The metaverse's popularity could be imposed

A predominant framework for assessing any technology's attractiveness for malevolent actors relies on four key aspects: audience reach, stability, usability, and platform security (Tech Against Terrorism 2021). As of now, we believe that the malevolent exploitation of the metaverse is perceived as relatively insignificant based on its slow rise in popularity together with persistent problems of popular social media platforms that continue to dominate headlines—especially since Elon Musk's Twitter takeover (Spring 2023). However, that view overwhelmingly relies on the fact that the metaverse reaches far fewer people compared to "traditional" social media platforms as justification for a lack of impact (Wang et al. 2022). A glaring fact remains—the constituent technologies of XR are being pushed commercially, and are continuously invested in, by all of the leading tech firms. Therefore, this technology differentiates itself from other, similarly fringe, technologies that are not currently exploited by malevolent actors on a large scale (Trauthig and Bodo 2022), because the metaverse is likely to grow and continue to attract attention. Both of these features are likely to increase its attractiveness for bad actors hoping to cajole, harass, and exploit.

The popularity of a technology does not only grow organically in a bottom-up, consumer-driven, way. Authoritarian countries like China have proven that technologies can massively scale without consumer buy-in due to top-down rollout of technology (Parasol 2022). We may witness similar trends with regard to the metaverse, both in China and beyond. The ruling Communist Party in China (CCP) has already relied on VR technologies to test and solidify party loyalty. For instance, in Shandong province in the East of China, party members have been required to participate in political loyalty tests that employed VR. According to a popular industry blog called VR Focus, the "Test of Dangxing," or "Test of Party Spirit," was obligatory for party members in Qingyang

(Hills-Duty 2018). They were asked to put on VR headsets and join a virtual room, in which leading figures of the party quizzed them on a multitude of subjects, such as party theory and how they understood the "pioneering role" of the CCP, but also inquired about members' daily lives (Hills-Duty 2018).

What are the CCP's reasons for relying on a VR experience instead of a phone or video conference call, in either of which they could have asked the same questions? First, VR is more immersive, and therefore more physiologically potent. It is, even, more threatening because people's learned threat responses have largely been trained in real, not virtual, environments over the course of their lifetimes (Rosén et al. 2019). The consequences of performing badly, or behaving dishonestly, feel more real in a multisensory environment. According to the group that administered the VR experiences, the test results were used to inform and identify the people and particular qualities that the CCP wants to promote. VR, in short, can be a superior medium for manipulating others—not only through mis- or disinformation but also through sensory control (Hills-Duty 2018). Given that China accounted for 82% of global shipments for VR headsets in 2019, these developments are likely to continue (Baptista 2019).

Potentially, nondemocratic countries could have a significant impact on the success of the metaverse based on user uptake but also on political manipulation. Democratic backsliding or advanced authoritarianism defines the majority of the five most populous countries in the world: China, India, the US, Indonesia, and Pakistan (with the Bertelsmann Transformation Index designating China and Pakistan as hardline autocracies, and India and Indonesia as defective democracies) (Bertelsmann Foundation 2022). Since China in particular has been implementing policy initiatives (such as the Chinaverse initiative from 2021) for the metaverse that are defined by state control of all its aspects (Ball 2022), and other authoritarian countries like the United Arab Emirates have been investing heavily in the metaverse (Kshetri 2023), politically motivated manipulation of public opinion that could include the metaverse may be unlikely to play out in the US alone (Huang and Purnell 2023).

2.2 The metaverse poses new questions about negative impacts on the individual user

The varying impact of manipulative content spread over different social media platforms and other digital spaces has been much discussed and studied over the last several years (Mattingly and Yao 2022). Some have derided disinformation spread over social media as being less-than-impactful and/or (perhaps adding to a lack of understanding about effect) difficult to define and study systematically (Tucker et al. 2018). Worryingly, however, existing studies of political propaganda spread over private messaging apps, such as Telegram, Signal, or WhatsApp,offer initial data points suggesting that manipulative content on closed platforms can have stronger impacts than similar content spread over Facebook or Twitter (Machado et al. 2019). People may tend to have their guard down on platforms that they associate with communication with close friends and family. They hence see information they receive via these platforms as more authentic and take it more seriously than information on their Facebook timeline, for instance (Trauthig and Mimizuka 2022). This is related to a mechanism referred to as "relational organizing"—in this case, a marketing framework by which propagandists and other manipulative actors purposefully exploit close ties (Gursky et al. 2022).

The metaverse adds renewed emphasis on these deliberations about potency and impact of manipulative content. Virtual reality and the advanced technologies related to it hold the potential to be deployed effectively to spread computational propaganda: automation and AI can be used over XR to manipulate public opinion in ways that involve less direct persuasion of users and more gaming of recommendation algorithms (Woolley 2020). Simultaneously, XR allows for multisensory disinformation that "immerses" in falsehood to be much more difficult to ignore than sidebar ads on Facebook (Woolley 2020). Given this, informational manipulation over the metaverse has the potential to be particularly potent because this suite of technologies places users in a digitally advanced landscape where they not only see but also hear and feel the ecosystem around them. In these environments people are more likely to emotionally relate as they are quite literally put in the midst of multisensory disinformation. The science detailing these psychological effects is still progressing, but initial results show they are powerful (Han, Bergs, and Moorhouse 2022).

The added potency of propaganda in the metaverse is also related to difficulties in identifying it. VR expert Toshi Anders Hoo explains that while users might be capable of detecting signs of manipulation in some digital media—like seeing altered blinking patterns or funny-looking lip movements in a deepfake or hearing improper modulation of an automated voice in a video—they will have significantly more of a challenge in catching such tells when the two senses are merged with movements in an immersive media environment (Woolley 2020). How, Anders Hoo asks, do we feel fakeness?

Finally, these effects are likely to impact various age groups differently. Existing VR businesses tailored toward a range of age groups are, themselves, aware of potential bad effects for their users. For example, Joseph Sullivan, founder of the California-based VR firm Luciton Virtual, emphasized that people over retirement age are particularly vulnerable to VR-driven propaganda (Newton 2019). As a consequence, Luciton Virtual communicates to customers that VR experiences are an insufficient replacement for real human interaction and highlights that, just as in real life, bad actors might try to take advantage of people in VR experiences. With VR being adopted as an educational technology, children's engagements with the technology is also cause for concern and continued monitoring. While experiences in the virtual classroom might be supervised by educators during class time, this introduction to VR can proliferate into increased time spent in virtual environments that are less supervised. As 2022 observed in digital ethnographic work, "children have to bear the social pressure of trying to fit in and emulate adults' behaviours such as...sexual activities in VRChat," and those children "who are new to the platform could find themselves in even riskier situations due to the higher chance of encountering harassing behaviours in the public domain." (Hu (2022) differentiates between three different layers of VRchat's social worlds: public, semiprivate, and private).

Given the studies that researched underage grooming and children's inappropriate exposure to sexual imagery in the metaverse (Hu 2022), the focus on differentiating impact of a wide range of types of propaganda in the metaverse by sociological criteria such as age—but also race—is valuable. By 2023, a plethora of studies have explained how one-to-many communication technologies like television affect both young and elderly viewers, but we know significantly less about the impact of many-to-many digital tools. Analyzing possible smartphone or video game "addictions" is relatively new by research standards but has important implications for a potential future filled with immersive experiences (Das et al. 2017). How long will it be before we have a meaningful understanding of the effect of XR media on the human brain? Upon the developing brains of young people? Almost by definition it will be impossible to examine large-scale data until long after those technologies have been released and widely adopted (Woolley 2020).

2.3 Users have poor protection from hateful or misleading content in the metaverse

The increased difficulty of creating safe environments in the metaverse (BBC News 2023; Bajwa 2022) can be a harbinger for increased malevolent exploitation for reasons very much related to the prior framework for assessing any technology's attractiveness for malevolent actors (based on audience reach, stability, usability, and security of platforms).

Child exploitation, disenfranchisement, harassment of marginalized groups, and coordinated influence operations could spread more easily over VR worlds due to increased difficulties in moderating them and presumably lacklustre investments of respective companies. In order to test Meta's efforts at content moderation and reactions to users' exposure to hateful and misleading content, reporters from BuzzFeed built their own private world inside Horizon (Baker-White 2022). Their "Qniverse" tested the company's VR-moderation systems, including detection, reporting mechanisms, and enforcement, over the course of a few days (Baker-White 2022). The team flooded their virtual world with egregious disinformation; for example, they played a relentless soundtrack of conspiracy theorist Alex Jones calling US president Joe Biden a pedophile. They broadcast the infamous "Stop the Steal!" rhetoric favored by previous US president Trump as well as the "Stop the Plandemic!" slogan, which drove a major COVID-19-related disinformation campaign (Neuman 2020).

In other words, this metaverse's sky was covered in text that Meta has explicitly promised to remove from its other social media platforms, particularly Facebook and Instagram (Meta 2020). Eventually, one of the reporters used Horizon's user reporting function (relying on a linked Facebook account) to notify Meta of the Qniverse. For two days, there was no reaction at all, so the same user reported the world again, and another reporter relied on her Facebook and Oculus profiles to do the same. Then Meta responded that their "trained safety specialist" reviewed the report and "determined that the content in the 'Qniverse' doesn't violate our Content in VR Policy" ((Baker-White 2022)). Meta's decision-making process in this instance is unclear—potentially the moderators decided not to act because the world "only" contained abusive and misleading speech but no behavior. Or perhaps the experimental world was considered too small to be of any significance. Surprisingly, however, 24 hours later, the Qniverse had disappeared as the company took a U-turn on its initial decision. This virtual reality in Horizon Worlds was purposefully reported for suspicious activity by its own builders and members due to the nature of the world (as an experiment of reporters). Other "unpublished" worlds created by incubators of extremist ideas or child exploitation are unlikely to be reported from the inside, given that most members would hate to see their hidden virtual realities disappear (Baker-White 2022).

Thus, there is accumulating evidence that content moderation systems that operate on Facebook and Instagram do not function on Horizon. Moreover, in virtual reality, harmful content or behavior like trolling can be more potent and visceral than what users might endure on other social media platforms (McCarthy 2022). However, one Meta employee who works as an engineer at Reality Labs told our research team that he does not believe "anyone thinks about those things...[because] we have bigger fish to fry right now."

3 Call for action

Are private companies and users operating in the metaverse ready to contend with malevolent exploitation of the space? We think no. There are three reasons why this is the case.

3.1 Existing company structures are not designed for prioritizing user safety

One concern is that companies prioritize profit over safety. For example, Meta is a profit-seeking company under financial strain. It has disappointed consumers, shareholders, and policymakers by putting profit over user safety in the past (Jurecic and Rozenshtein 2021). There are still unanswered issues with Facebook and Instagram algorithms (Allcott, Gentzkow, and Yu 2019). It is unclear, however, if Meta's algorithms can amplify virtual worlds in the same way they have boosted groups that spread conspiracy theories, electoral disinformation, and digital hate; in fact, Meta declined to answer when asked once about whether its recommendations guidelines and/or content distribution guidelines apply in VR (Baker-White 2022). And while the company's vice president of VR retroactively declared that his 2016 statement asserting that Facebook needs to prioritize growth over curbing offline harms like terrorism was deliberately provocative (Mac, Warzel, and Kantrowitz 2018), unfortunately, we believe that Meta's bottom line still remains: there is profit to be had from creating division and allowing hate (BBC News 2020), while ensuring online safety is a pricey endeavor (Alba and Wagner 2023).

The second main concern is the well-documented ability of malevolent actors to innovate and prey on systematic vulnerabilities of online platforms (Trauthig and Bodo 2022; Bajwa 2022). In short, propagandists and extremists are always looking for new ways to share content while evading detection. And malevolent actors deliberately working toward finding and exploiting emerging systems' weakest points of enforcement are already becoming interested in the metaverse—the more people they can reach, radicalize, or recruit (Bajwa 2022). We are unsure that existing company structures of firms invested in the XR/VR space are agile enough to follow or ideally anticipate future malign behavior.

The third and final concern is that existing journalistic inquiries and outside research projects have pointed toward an unclear and presumably resource-scarce effort of companies to deal with complaints of misleading and hateful content and behavior in the metaverse (Berryman and Leaver 2022; Baker-White 2022) —and companies are likely to struggle more with the possible increased success of the metaverse. Currently, on Horizon Worlds for example, there are four ways a user can protect themselves from misleading and hateful content or behavior: report an individual user or a world to Meta for violating its Code of Conduct for Virtual Experiences, remove oneself into a personal safe space ("safe zone"), block individual users from your safe zone, and create personal boundaries that surround the user's avatar.[1] We think it likely that the first way (reporting), especially, will pick up if more users join Horizon Worlds, and we are skeptical that companies, like Meta for example, will invest corresponding additional resources into teams dealing with user reports.

However, as was mentioned in the introduction, not everyone is aware of those options and, similar to issues users experience with reporting harmful content or behavior on Facebook, delays and lack of clarity sometime accompany the reporting process (Gillespie 2018). In the ratings and review section of Horizon Worlds on Meta's homepage, eight out of the ten user reviews rated most relevant complain about abusive

[1] https://www.meta.com/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds

speech or behavior they encountered or unclear user removals they suffered (the other two emphasize technical problems when entering Horizon Worlds).[2] While the number of Horizon users may currently be small enough for Meta to enable this type of reactive moderation—such as user removal after reporting and follow-up on all of the reports by having staff or contractors monitor everything that users flag—they are likely to reach further limits as the platform grows. Similarly, companies like AltSpaceVR, which is part of Microsoft's mixed-reality division, seem to have been following a similar approach of trying to put moderation tools into the hands of its users with occasional staff support (Mak 2022). This approach is largely related to declared privacy issues with recording everything that happens and proactively intervening (Bosworth 2021). The crucial question here is if companies invested into VR experiences put sufficient resources into teams that proactively factor in and develop detection of malevolent content and behavior. We need only remember how quickly Facebook went from a platform used by millions of people to one used by billions—and, correspondingly, how quickly the platform's informational woes became unmanageable due to sheer scale (Vaidhyanathan 2018)—to see that more systematic, scalable action related to trust and safety in the metaverse should happen now.

And what about the dangers that can emanate from private virtual worlds like the one described in section 2.3? What about communities filled with harmful content that might never be reported but that serve as incubators for hate, harassment, and propaganda? Again, this concern is related to the alleged focus on reactive moderation instead of proactive detection (Medina et al. 2022), which we, as outsiders, are unable to fully understand. Relying on users, or volunteer or ad-hoc recruited moderators, in the metaverse would put the onus on people who are neither prepared nor solely responsible for what awaits them. Aaron Mak, a former Slate Tech reporter, spoke to different VR moderators—all had "horror stories about dealing with troublemakers" in the space (Mak 2022). The metaverse is already interconnected with the gaming world, which has a history of toxic behaviors (Hoang 2021). Will this continue to seep into other, non-gamified XR spaces? For instance, Roblox, another metaverse frontrunner particularly popular among kids, has a staggering two billion player accounts. It also regularly makes headlines due to the proliferation of disturbing user content (D'Anastasio 2021). As Heller argues, "technology to adequately moderate in VR doesn't exist yet" and Big Tech's approaches so far seem inadequate (Atlantic Council 2022).

3.2 Recommendations

First, companies, including game and headset creators that are part of the XR Association,[3] should adopt and follow transparency protocols related to their reporting tools and resulting actions. Introducing these mechanisms should be part of the onboarding process after users buy a new device. Many platforms provide tools to tackle harassment and abuse, but few people know how to use them or trust they will work (Phippen 2022). Transparency around these rules—and efficacy of related procedures—will be vital as Meta and others continue to work to provide digital environments free of hateful and misleading content/behavior. A good example is the European Digital Services Act, which requires new transparency compliance, through reporting obligations for companies' terms of service or audited reports about content moderation. Meta regularly pledges transparency around its decision-making in this regard, but when approached for comments about how it protects people in its XR, concrete answers have not been given. This was the experience of the BuzzFeed News

[2] https://www.oculus.com/experiences/quest/2532035600194083/

[3] https://xra.org/

Team referenced earlier, who instead received an elusive, broad statement from a Meta spokesperson emphasizing that Meta remains guided by their "Responsible Innovation Principles" (Baker-White 2022).

Another clear step toward accountability that Meta and other XR firms should take is to support the creation and maintenance of a third-party commission aimed at administering the difficult analyses and decisions needed in the face of the current problems with the digital information ecosystem. Critically, this commission should not be intrinsically tied to a specific firm, as the Oversight Board is to Meta, but have an international, multiplatform mandate. Experts have argued for the creation of just such a governance body—an "International Panel on the Information Environment" modeled on the Intergovernmental Panel on Climate Change. Such a panel would "empirically monitor, verify, and inspect the actions of technology firms and government agencies" (Himelfarb and Howard 2022).

Second, policymakers should consider the metaverse a future that has already arrived. Given this, they need to incorporate it into ongoing discussions about tech regulation. In some instances, especially in the European Union, this is already happening, but in others tech lobbyists seem to be succeeding in convincing policymakers of the alleged futility when trying to regulate something as "fresh and new" as the metaverse (Dwoskin, Zakrzewski, and Miroff 2021). The US, in particular, stands out as a country that has continuously failed to sensibly regulate the online sphere, and especially social media and newer spaces like the metaverse. At the most basic level, the US needs privacy laws to prevent the buying and selling of peoples' online data. That data is, after all, a core means of bespoke manipulation used by propagandists and professional trolls as they work to target particular groups and subgroups.

Transparency requirements related to trust and safety standards and enforcement contain many sensitive questions, but these requirements should be in focus due to their importance for a well-informed society and, we believe, should take precedence over calls for implementing accountability for companies' proprietary algorithms, for example. Existing standards—such as checking uploaded files for child exploitation and terrorist material against hashed-image banks maintained by organizations like the National Center for Missing and Exploited Children and the Global Internet Forum to Counter Terrorism (GIFCT)—should be transferred to VR/XR realities. Updating existing processes, such as image sharing, to equally useful mechanisms for VR worlds is an important task and can hopefully build on existing collaborations. Potentially these processes are already under way, but from an outsider's perspective that does not seem to be the case as, e.g., the members' list of GIFCT indicates.[4] In other words, regulatory bodies should incentivize companies to disclose how they mitigate challenges in the metaverse; how they protect individuals from disinformation, harassment, or other cyberthreats; and with which resources. Similarly, existing regulations such as the European General Data Protection Regulation (GDPR), which is said to indirectly also apply to the metaverse, should be enforced (Robertson 2023), and American legislators should finally work on a similar regulatory framework (as referenced above).

With regard to Meta and the other firms currently dominating investment in VR and XR, enforcement antitrust legislation is particularly important to avoid them dominating the metaverse and therefore de facto set standards and avoid scrutiny via market domination. Decisions such as the one by the Federal Trade Commission to block Meta from acquiring Within Unlimited (a burgeoning VR contentmaker as fitness app) are steps in the right direction (FTC 2022), although the FTC received pushback when a California judge rejected the blocking of the acquisition of Within shortly after (Reuters

[4] https://gifct.org/membership/

2023).

Adequate regulations with regard to the metaverse could be approached in two different ways: First, policymakers could explore specific regulation for the metaverse in the form of a novel US Metaverse Act that would aim to incorporate all metaverse transactions and actions. While this would require efforts similar to the European Digital Services Act (DSA), the German NetzDG, or the planned Online Safety Bill drafted by UK legislators, the success of such legislation is in doubt due to the continued commitment and crosspartisan cooperation it would require. Second, regulators should incorporate and increase governmental oversight into the metaverse by tying metaverse legislation to straddling issue areas, such as copyright questions. As the metaverse includes some of the biggest technology firms (such as Meta and Microsoft) but also the more video game-focused companies (such as Roblox or Valve), ways to enhance the regulatory environment in the US include "state and federal gambling laws, money transfer laws, securities laws, and regulation of unfair and deceptive trade practices used to enforce privacy and cybersecurity obligations" (Garon 2022).

Third, the protection from hateful and misleading content and behavior in the metaverse requires a multistakeholder approach. Keeping individuals safe online and mitigating harmful processes like radicalization requires online and offline counteractions as well as cooperation from a variety of individuals and organizations. As a first step, policymakers, researchers, advocacy organizations, and corporations working on the metaverse should find regular fora for exchanging viewpoints. For example, third-party social scientists are essential to evaluate how the design of virtual worlds can affect societies and power relations within them. User education about online harms—and, as such, metaverse harms—needs to start early and should be embedded in a support system for tackling potential online harms, where no question is too dumb to be asked. Design of those programs should be led by the user perspective with different didactic approaches for different age groups, for example. Independent, community-based organizations could lead such educative efforts with financial backing from the companies selling VR devices.

Companies should continue to cooperate with each other but also with law enforcement to share information that will reduce risks and prevent content such as electoral disinformation from spreading or extremists from recruiting in the metaverse. Interpol, for example, has started looking into how the organization can investigate crime in the metaverse (BBC News 2023). Companies should also update their codes of ethics and whistleblowing programs to protect whistleblowers with regard to the metaverse, in case existing policies would not apply. Governments have a serious role in these protections, too, and must consider them. The far-reaching consequences an internationally successful metaverse might have for societies around the globe brings up the need for an international and global authority to oversee the metaverse. Given the established structure of the United Nations, it would be sensible to attach this authority there; however, the main question would be how to make it a body with teeth independent of the political whims of the Security Council.

Fourth, those involved in developing the metaverse should be embedded in an ethically focused design framework that acknowledges that societal inequalities often morph into online marginalization. Harassment of minority communities and women has been a problem that Horizon and other VR worlds have struggled with from inception (Basu 2021). Ensuring that this next iteration of the internet is inclusive and works for everyone will require that people from marginalized communities not only are factored in from the start but also take lead positions in designing it (Costanza-Chock 2020). While utopian visions in the early days of the internet were abound with promoting a radically different, better, nondiscriminatory experience of online life (Winner 1997), the internet turned out to be far from raceless (Nakamura 2013). In short, the metaverse needs design justice that puts people who have little power in society at the center of the design process to circumvent the perpetuation of existing inequalities. It also includes initiating the deliberation of values and principles to guide design (Adeyemo 2021).

For these recommendations to be implemented, there must be a joint approach across sectors. Lawmakers should work together across party lines—and ideally across the Atlantic and the world. The European Union is continuing with its proactive approach to tech regulation and has started planning a full policy initiative on the metaverse for 2023 (Robertson 2023). While these are challenging demands, online trust and safety should be important to everyone—especially considering the fact that many children and vulnerable communities use the internet and XR. Given large investments in the space, "the metaverse" is likely to simply morph into what future users will consider to be "the internet." We should not allow this slow move toward XR as an ambient technology: another tool that simply coalesces into the digital world we already understand.

References

Adeyemo, Breigha. 2021. “I’m a Black woman and the metaverse scares me – here’s how to make the next iteration of the internet inclusive.” *The Conversation* (December). http://theconversation.com/im-a-black-woman-and-the-metaverse-scares-me-heres-how-to-make-the-next-iteration-of-the-internet-inclusive-173310.

Ahn, Sun Joo (Grace), Jeremy N. Bailenson, and Dooyeon Park. 2014. “Short- and long-term effects of embodied experiences in immersive virtual environments on environmental locus of control and behavior.” *Computers in Human Behavior* 39 (October): 235–45. Accessed February 15, 2023. https://www.sciencedirect.com/science/article/pii/S0747563214003999.

Alba, Davey, and Kurt Wagner. 2023. “Twitter’s Trust and Safety Head Ditches Protocol or Musk Whims.” *Bloomberg* (January). https://www.bloomberg.com/news/articles/2023-01-27/elon-musk-s-twitter-trust-safety-head-ella-irwin-breaks-rules-for-him.

Alfonseca, Kiara, and Max Zahn. 2023. “Tech layoffs 2023: Companies that have made cuts.” *ABC News* (February). Accessed February 14, 2023. https://abcnews.go.com/Business/tech-layoffs-2023-companies-made-cuts/story?id=96564792.

Allcott, Hunt, Matthew Gentzkow, and Chuan Yu. 2019. “Trends in the diffusion of misinformation on social media.” *Research & Politics* 6, no. 2 (April). Accessed February 14, 2023. https://doi.org/10.1177/2053168019848554.

Atlantic Council. 2022. “What happens when toxic online behavior enters the metaverse?” June. Accessed January 23, 2023. https://www.atlanticcouncil.org/news/transcripts/what-happens-when-toxic-online-behavior-enters-the-metaverse/.

Atske, Sara. 2022. “The Metaverse in 2040.” *Pew Research Center: Internet, Science & Tech* (June). Accessed January 23, 2023. https://www.pewresearch.org/internet/2022/06/30/the-metaverse-in-2040/.

Bailenson, Jeremy. 2018. *Experience on Demand: What Virtual Reality Is, How It Works, and What It Can Do.* W. W. Norton & Company, January.

Bajwa, Aman. 2022. “Malevolent Creativity & the Metaverse: How the immersive properties of the metaverse may facilitate the spread of a mass shooter’s culture.” *The Journal of Intelligence, Conflict, and Warfare* 5, no. 2 (November): 32–52. Accessed February 2, 2023. https://journals.lib.sfu.ca/index.php/jicw/article/view/5038.

Baker-White, Emily. 2022. “Meta Wouldn’t Tell Us How It Enforces Its Rules In VR, So We Ran A Test To Find Out.” *BuzzFeed News* (February). Accessed January 30, 2023. https://www.buzzfeednews.com/article/emilybakerwhite/meta-facebook-horizon-vr-content-rules-test.

Ball, Matthew. 2022. *The Metaverse: and How it Will Revolutionize Everything.* Liveright Publishing.

Baptista, Eduardo. 2019. “China’s Communist Party Is Making Its Own (Virtual) Reality.” *Foreign Policy* (June). Accessed February 2, 2023. https://foreignpolicy.com/2019/06/21/chinas-communist-party-is-making-its-own-virtual-reality/.

Basu, Tanya. 2021. “The metaverse has a groping problem already.” *MIT Technology Review* (December). Accessed February 2, 2023. https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/.

BBC News. 2020. "Facebook 'profits from hate' claims engineer who quit." *BBC News* (September). Accessed February 1, 2023. https://www.bbc.com/news/technology-54086598.

———. 2023. "Interpol working out how to police the metaverse." *BBC News* (February). Accessed February 6, 2023. https://www.bbc.com/news/technology-64501726.

Berryman, Rachel, and Tama Leaver. 2022. "'Virtual influencers' are here, but should Meta really be setting the ethical ground rules?" *The Conversation* (January). Accessed January 23, 2023. http://theconversation.com/virtual-influencers-are-here-but-should-meta-really-be-setting-the-ethical-ground-rules-175524.

Bertelsmann Foundation. 2022. *Bertelsmann Transformation Index 2022: Political Transformation* [in en]. Technical report. Accessed February 27, 2023. https://bti-project.org/en/index/political-transformation.

Bosworth, Andrew. 2021. "Keeping People Safe in VR and Beyond." *Meta Quest Blog* (November). Accessed February 1, 2023. https://www.oculus.com/blog/keeping-people-safe-in-vr-and-beyond/.

Casswell, Sally. 2022. "Alcohol marketing has crossed borders and entered the metaverse – how do we regulate the new digital risk?" *The Conversation* (May). Accessed January 23, 2023. http://theconversation.com/alcohol-marketing-has-crossed-borders-and-entered-the-metaverse-how-do-we-regulate-the-new-digital-risk-183334.

Chen, Chuan, Lei Zhang, Yihao Li, Tianchi Liao, Siran Zhao, Zibin Zheng, Huawei Huang, and Jiajing Wu. 2022. "When Digital Economy Meets Web3.0: Applications and Challenges." *IEEE Open Journal of the Computer Society* 3:233–45. https://doi.org/10.1109/OJCS.2022.3217565.

Chohan, Usman W. 2022. "Metaverse or Metacurse?" *SSRN Electronic Journal,* accessed February 27, 2023. https://www.ssrn.com/abstract=4038770.

Costanza-Chock, Sasha. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need.* Cambridge, Massachusetts: The MIT Press, March.

D'Anastasio, Cecilia. 2021. "How 'Roblox' Became a Playground for Virtual Fascists." *Wired* (June). Accessed February 1, 2023. https://www.wired.com/story/roblox-online-games-irl-fascism-roman-empire/.

Daniel, Will. 2022. "Meta's metaverse business is losing billions, but Mark Zuckerberg says it's all part of the plan." *Fortune* (April). Accessed January 23, 2023. https://fortune.com/2022/04/28/meta-facebook-mark-zuckerberg-metaverse-business-losing-billions-part-plan/.

Das, Prithwijit, Zhu Meng'ou, Laura McLaughlin, Zaid Bilgrami, and Ruth Milanaik. 2017. "Augmented Reality Video Games: New Possibilities and Implications for Children and Adolescents." *MDPI* (April). https://doi.org/10.3390/mti1020008.

Dionisio, John David N., William G. Burns III, and Richard Gilbert. 2013. "3D Virtual worlds and the metaverse: Current status and future possibilities." *ACM Computing Surveys* 45, no. 3 (July). Accessed January 23, 2023. https://doi.org/10.1145/2480741.2480751.

Doctor, Austin C., Joel S. Elson, and Sam Hunter. 2022. "The metaverse offers a future full of potential – for terrorists and extremists, too." *The Conversation* (January). Accessed January 23, 2023. http://theconversation.com/the-metaverse-offers-a-future-full-of-potential-for-terrorists-and-extremists-too-173622.

Dwivedi, Yogesh K., Laurie Hughes, Abdullah M. Baabdullah, Samuel Ribeiro-Navarrete, Mihalis Giannakis, Mutaz M. Al-Debei, Denis Dennehy, et al. 2022. “Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy.” *International Journal of Information Management* 66 (October). Accessed January 23, 2023. https://doi.org/10.1016/j.ijinfomgt.2022.102542.

Dwoskin, Elizabeth, Cat Zakrzewski, and Nick Miroff. 2021. “How Facebook’s ‘metaverse’ became a political strategy in Washington.” *Washington Post* (September). Accessed March 1, 2023. https://www.washingtonpost.com/technology/2021/09/24/facebook-washington-strategy-metaverse/.

Federal Trade Commission. 2022. “FTC Seeks to Block Virtual Reality Giant Meta’s Acquisition of Popular App Creator Within” (July). Accessed February 2, 2023. https://www.ftc.gov/news-events/news/press-releases/2022/07/ftc-seeks-block-virtual-reality-giant-metas-acquisition-popular-app-creator-within.

Garon, Jon. 2022. “Legal Implications of a Ubiquitous Metaverse and a Web3 Future,” January. SSRN Scholarly Paper. Accessed January 23, 2023. https://doi.org/10.2139/ssrn.4002551.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media.* Yale University Press.

Gursky, Jacob, Martin J. Riedl, Katie Joseff, and Samuel Woolley. 2022. “Chat Apps and Cascade Logic: A Multi-Platform Perspective on India, Mexico, and the United States.” *Social Media + Society* 8, no. 2 (April). https://doi.org/10.1177/20563051221094773.

Han, Dai-In Danny, Yoy Bergs, and Natasha Moorhouse. 2022. “Virtual reality consumer experience escapes: preparing for the metaverse.” *Virtual Reality* 26, no. 4 (December): 1443–58. Accessed February 2, 2023. https://doi.org/10.1007/s10055-022-00641-7.

Heller, Brittan, and Avi Bar-Zeev. 2021. “The Problems with Immersive Advertising: In AR/VR, Nobody Knows You Are an Ad.” *Journal of Online Trust and Safety* 1, no. 1 (October). Accessed February 2, 2023. https://tsjournal.org/index.php/jots/article/view/21.

Hills-Duty, Rebecca. 2018. “China’s Communist Party Uses VR for Loyalty Test” [in en]. *VR Focus* (May). Accessed February 2, 2023. www.vrfocus.com/2018/05/chinas-communist-party-uses-vr-for-loyalty-tests/.

Himelfarb, Sheldon, and Phil Howard. 2022. “Global information wars are a threat to the planet. This is how we could broker peace.” *BostonGlobe.com,* accessed March 8, 2023. https://www.bostonglobe.com/2022/07/07/opinion/global-information-wars-are-threat-planet-this-is-how-we-could-broker-peace/.

Hoang, Ly. 2021. “A new survey of gamers shows toxicity, misogyny, and a lack of effective moderation risks becoming normalized.” *Utopia Analytics* (November). Accessed February 1, 2023. https://utopiaanalytics.com/a-new-survey-of-gamers-shows-toxicity-misogyny-and-a-lack-of-effective-moderation-risks-becoming-normalized/.

Hu, Runze. 2022. “Understanding children’s vulnerabilities in the metaverse: the role of the online community.” *London School of Economics and Political Science* (June). Accessed February 14, 2023. https://blogs.lse.ac.uk/parenting4digitalfuture/2022/06/15/metaverse-vrchat/.

Huang, Raffaele, and Newley Purnell. 2023. "Meta in Talks to Reboot China Business With VR Headsets." *Wall Street Journal,* https://www.wsj.com/articles/tencent-in-talks-to-sell-metas-quest-2-vr-headset-in-china-6d90190a.

Jurecic, Quinta, and Alan Rozenshtein. 2021. "Mark Zuckerberg's Metaverse Unlocks a New World of Content Moderation Chaos." *Lawfare* (November). Accessed January 23, 2023. https://www.lawfareblog.com/mark-zuckerbergs-metaverse-unlocks-new-world-content-moderation-chaos.

Kirkpatrick, Keith. 2022. "Applying the metaverse." *Communications of the ACM* 65, no. 11 (November): 16–18. Accessed February 27, 2023. https://dl.acm.org/doi/10.1145/3565470.

Kshetri, Nir. 2023. "National metaverse strategies." *Computer* 56 (2): 137–42.

Louise, Nickie. 2022. "The Dark Side Of Metaverse—Experts warn Facebook's metaverse poses 'terrifying dangers.' Facebook's VR Metaverse (VRChat) has become a cesspool for predators who prey on children" (January). Accessed January 23, 2023. https://techstartups.com/2022/01/25/dark-side-metaverse-experts-warn-facebooks-metaverse-poses-terrifying-dangers-everyone-facebooks-vr-metaverse-vrchat-become-cesspool-predators/.

Ma, Adrian. 2022. "What is the metaverse, and what can we do there?" *The Conversation* (May). Accessed January 23, 2023. http://theconversation.com/what-is-the-metaverse-and-what-can-we-do-there-179200.

Mac, Ryan, Sheera Frenkel, and Kevin Roose. 2022. "Skepticism, Confusion, Frustration: Inside Mark Zuckerberg's Metaverse Struggles." *The New York Times* (October). Accessed October 12, 2022. https://www.nytimes.com/2022/10/09/technology/meta-zuckerberg-metaverse.html.

Mac, Ryan, Charlie Warzel, and Alex Kantrowitz. 2018. "Top Facebook Executive Defended Data Collection In 2016 Memo — And Warned That Facebook Could Get People Killed." *BuzzFeed News,* accessed February 1, 2023. https://www.buzzfeednews.com/article/ryanmac/growth-at-any-cost-top-facebook-executive-defended-data.

Machado, Caio, Beatriz Kira, Vidya Narayanan, Bence Kollanyi, and Philip Howard. 2019. "A Study of Misinformation in WhatsApp groups with a Focus on the Brazilian Presidential Elections." In *Companion Proceedings of The 2019 World Wide Web Conference,* 1013–19. WWW '19. New York, NY, USA: Association for Computing Machinery, May. Accessed February 2, 2023. https://doi.org/10.1145/3308560.3316738.

Mak, Aaron. 2022. "I Was a Bouncer in the Metaverse." *Slate* (May). Accessed January 23, 2023. https://slate.com/technology/2022/05/metaverse-content-moderation-virtual-reality-bouncers.html.

Mattingly, Daniel C., and Elaine Yao. 2022. "How Soft Propaganda Persuades." *Comparative Political Studies* 55, no. 9 (August): 1569–94. Accessed February 2, 2023. https://doi.org/10.1177/00104140211047403.

McCarthy, Dan. 2022. "A new challenge for Meta: How to moderate the metaverse." *Emerging Tech Brew* (February). Accessed January 23, 2023. https://www.emergingtechbrew.com/stories/2022/02/23/a-new-challenge-for-meta-how-to-moderate-the-metaverse.

McEwan, Bree. 2021. “We know better than to allow Facebook to control the metaverse.” *The Conversation* (November). Accessed January 23, 2023. http://theconversation.com/we-know-better-than-to-allow-facebook-to-control-the-metaverse-171467.

Medina, Robin, Judith Njoku, Jae Min Lee, and Dong-Seong Kim. 2022. “Audio-Based Hate Speech Detection for the Metaverse using CNN.” November.

Meta. 2020. “An Update to How We Address Movements and Organizations Tied to Violence,” August. Accessed January 30, 2023. https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/.

———. 2021. “Connect 2021: Our vision for the metaverse,” October. Accessed January 23, 2023. https://tech.facebook.com/reality-labs/2021/10/connect-2021-our-vision-for-the-metaverse/.

Meta Quest. 2022. “Code of Conduct for Virtual Experiences Meta Store,” November. Accessed February 27, 2023. https://www.meta.com/help/quest/articles/accounts/privacy-information-and-settings/code-of-conduct-for-virtual-experiences/?utm_source=developer.oculus.com&utm_medium=oculusredirect.

Nakamura, Lisa. 2013. *Cybertypes: Race, Ethnicity, and Identity on the Internet.* New York: Routledge, May.

Neuman, Scott. 2020. “Seen ‘Plandemic’? We Take A Close Look At The Viral Conspiracy Video’s Claims.” *NPR* (May). Accessed January 30, 2023. https://www.npr.org/2020/05/08/852451652/seen-plandemic-we-take-a-close-look-at-the-viral-conspiracy-video-s-claims.

Newton, Casey. 2019. “People Older Than 65 Share the Most Fake News, a New Study Finds.” *The Verge* (January). www.theverge.com/2019/1/9/18174631/old-people-fake-new-facebook-share-nyu-princeton.

Parasol, Max. 2022. “China’s Digital Yuan: Reining in Alipay and WeChat Pay.” *Banking & Finance Law Review* 37, no. 2 (April): 265–303. Accessed January 25, 2023. https://www.proquest.com/docview/2653586051/abstract/9DAAE19101B94247PQ/1.

Phippen, Andy. 2022. “Protecting children in the metaverse: it’s easy to blame big tech, but we all have a role to play.” *The Conversation* (February). Accessed January 23, 2023. http://theconversation.com/protecting-children-in-the-metaverse-its-easy-to-blame-big-tech-but-we-all-have-a-role-to-play-177789.

Ravenscraft, Eric. 2022. “What Is the Metaverse, Exactly?” *Wired* (April). Accessed January 23, 2023. https://www.wired.com/story/what-is-the-metaverse/.

Reuters. 2023. “U.S. judge denies FTC request to stop Meta from acquiring VR firm Within.” *Reuters* (February). Accessed February 14, 2023. https://www.reuters.com/legal/us-federal-judge-denies-us-ftc-request-stop-meta-acquiring-virtual-reality-2023-02-04/.

Robertson, Derek. 2023. “How to regulate a universe that doesn’t exist.” *POLITICO* (February). Accessed February 8, 2023. https://www.politico.com/newsletters/digital-future-daily/2023/02/08/how-to-regulate-a-universe-that-doesnt-exist-00081895.

Rosén, Jörgen, Granit Kastrati, Aksel Reppling, Klas Bergkvist, and Fredrik Åhs. 2019. "The effect of immersive virtual reality on proximal and conditioned threat." *Scientific Reports* 9, no. 1 (November). Accessed February 14, 2023. https://www.nature.com/articles/s41598-019-53971-z.

Rosenberg, Louis. 2022. "There are two kinds of Metaverse. Only one will inherit the Earth." *Big Think* (January). Accessed January 23, 2023. https://bigthink.com/the-future/metaverse-augmented-virtual-reality/.

Shankland, Stephen. 2023. "Why the ChatGPT AI Chatbot Is Blowing Everybody's Mind." *CNET* (February). Accessed February 14, 2023. https://www.cnet.com/tech/computing/why-the-chatgpt-ai-chatbot-is-blowing-everybodys-mind/.

Silberling, Amanda. 2021. "Meta's Horizon Worlds is available in the US and Canada for 18+ users." *TechCrunch* (December). Accessed February 27, 2023. https://techcrunch.com/2021/12/09/metas-horizon-worlds-is-available-in-the-us-and-canada-for-18-users/.

Slater, Mel, Maria V. Sanchez-Vives, Albert Rizzo, and Massimo Bergamasco, eds. 2019. *The Impact of Virtual and Augmented Reality on Individuals and Society.* Frontiers Media SA.

Smaili, Nadia, and Audrey de Rancourt-Raymond. 2022. "Metaverse: Welcome to the new fraud marketplace." *Journal of Financial Crime,* no. ahead-of-print, accessed January 23, 2023. https://doi.org/10.1108/JFC-06-2022-0124.

Spring, Marianna. 2023. "Twitter insiders: We can't protect users from trolling under Musk." *BBC* (March). https://www.bbc.com/news/technology-64804007.

Tech Against Terrorism. 2021. *Gap Analysis and Recommendations for deploying technical solutions to tackle the terrorist use of the internet.* Technical report. GIFCT, July. https://gifct.org/wp-content/uploads/2021/07/GIFCT-TAWG-2021.pdf.

Trauthig, Inga, and Lorand Bodo. 2022. *Emergent Technologies and Extremists: The DWeb as a New Internet Reality?* Technical report. Global Network on Extremism and Technology, August. Accessed January 23, 2023. https://gnet-research.org/2022/08/01/emergent-technologies-and-extremists-the-dweb-as-a-new-internet-reality/.

Trauthig, Inga, and Kayo Mimizuka. 2022. *WhatsApp, Misinformation, and Latino Political Discourse in the U.S.* Technical report. Tech Policy Press, October. Accessed February 2, 2023. https://techpolicy.press/whatsapp-misinformation-and-latino-political-discourse-in-the-u-s/.

Tucker, Joshua A., Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature.* Technical report. Hewlett Packard. Accessed February 14, 2023. https://hewlett.org/library/social-media-political-polarization-political-disinformation-review-scientific-literature/.

Vaidhyanathan, Siva. 2018. *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy.* Oxford University Press.

Wang, Yuntao, Zhou Su, Ning Zhang, Rui Xing, Dongxiao Liu, Tom H. Luan, and Xuemin Shen. 2022. "A Survey on Metaverse: Fundamentals, Security, and Privacy." *IEEE Communications Surveys & Tutorials,* https://doi.org/10.1109/COMST.2022.3202047.

Wiederhold, Brenda K. 2022. “Sexual Harassment in the Metaverse.” *Cyberpsychology, Behavior, and Social Networking* 25, no. 8 (August). Accessed January 23, 2023. https://doi.org/10.1089/cyber.2022.29253.editorial.

Winner, Langdon. 1997. “Technology Today: Utopia or Dystopia?” *Social Research* 64 (3): 989–1017. Accessed February 2, 2023. https://www.jstor.org/stable/40971195.

Woolley, Samuel. 2020. *The Reality Game: How the Next Wave of Technology Will Break the Truth.* New York: PublicAffairs, January.

Authors

Inga Kristina Trauthig is the head of research of the Propaganda Research Lab at the Center for Media Engagement based at The University of Texas at Austin. She is also an associate with the Institute of Middle Eastern Studies at King's College London (KCL) and previously has been a research fellow with the International Centre for the Study of Radicalisation at the Department of War Studies at KCL, where she received her PhD in Security Studies.

(inga.trauthig@austin.utexas.edu)

Samuel C. Woolley is an assistant professor at the School of Journalism and Media at The University of Texas at Austin and the program director of the propaganda research team and Knight faculty fellow at the Center for Media Engagement. He is an assistant professor (by courtesy) in UT's School of Information and a research affiliate at the Project for Democracy and the Internet at Stanford University.

Acknowledgements

The Propaganda Research Lab at the Center for Media Engagement at The University of Texas at Austin receives financial support by the John S. and James L. Knight Foundation, Miami Foundation, Omidyar Network, and Open Society Foundations.

Keywords

Metaverse, virtual reality, regulation, misinformation, radicalization.