ISD | Institute for Strategic Dialogue

Sponsored by:
Federal Foreign Office

Emerging Platforms and Technologies

An overview of the current threat landscape and its policy implications

Mauritius Dorn, Sara Bundtzen,
Christian Schwieter & Milan Gandhi

About the Digital Policy Lab

The Digital Policy Lab (DPL) is an inter-governmental working group focused on charting the policy path forward to prevent and counter the spread of disinformation, hate, extremist and terrorist content online. It comprises representatives of relevant ministries and regulatory bodies from liberal democracies. The DPL aims to foster inter-governmental exchange, provide policymakers and regulators with access to sector-leading expertise and research, and build an international community of practice around key challenges in the digital policy space. We thank the German Federal Foreign Office for their support of this project.

About this Paper

As part of the DPL, the Institute for Strategic Dialogue (ISD) organised two working group meetings on the topic of emerging platforms and technologies between May and June 2023. The working group consisted of DPL members representing ministries and regulatory bodies, including from Australia, Germany, New Zealand, Slovakia, the United Kingdom, the United States, and the European Union. Participants also included representatives from academia and civil society. While participants participated in the meetings and/or contributed to this publication, the views expressed in this policy paper do not necessarily reflect the views of all participants or any governments involved in this project.

Authors

Mauritius Dorn is a Senior Digital Policy and Education Manager at ISD Germany. He leads Project AHEAD – a dialogue series to provide an integrated understanding of hybrid threats with a focus on disinformation. He also supports the Digital Policy Lab (DPL).

Sara Bundtzen is an Analyst at ISD Germany, where she studies the spread of information manipulation in multilingual online environments. As part of the Digital Policy Lab (DPL), Sara analyses policy pathways toward countering disinformation, hate, and extremism.

Christian Schwieter is an ISD Fellow and PhD candidate at the Department of Media Studies at Stockholm University. Until 2023, he was Project Manager at ISD Germany, leading the research project 'Countering Radicalisation in Right-wing Extremist Online Subcultures'.

Milan Gandhi is a Research Fellow (AI and Public Policy) at ISD and a Master's candidate at the University of Oxford. He is also the founder of Legal Forecast, a not-for-profit organisation exploring the intersection of law and new technologies.

Editorial responsibility

Huberta von Voss, Executive Director, ISD Germany & Henry Tuck, Head of Digital Policy, ISD

Acknowledgements

We would like to thank all working group participants, including experts from academia and civil society for their contributions. We would like to give special thanks to the speakers as well as contributors to this paper for providing valuable insights and feedback, including: Diederik Don (Europol), Dr Elena Gubenko (Federal Office of Justice, Germany), Adam Hadley (Tech Against Terrorism), Dominik Hammer (ISD Germany), Dr Oliver Marsh (CASM Technology), and Jenna Omassi (Ofcom).

ISD | Institute for Strategic Dialogue

www.isdgermany.org

Table of Contents

Glossary

ActivityPub is an open and decentralised network protocol. As an open protocol, it does not belong to one particular company and is not limited to particular products or platforms. It provides client-to-server and server-to-server APIs. ActivityPub is a standard for the Fediverse.

Application programming interface (API) refers to a software intermediary that allows two applications to communicate with each other. APIs have a huge range of uses, but in the context of this report, they allow developers to integrate services into their applications. As an intermediary, APIs also provide an additional layer of security by not allowing direct access to data, alongside logging, managing and controlling the volume and frequency of requests.

BitTorrent is a Peer-to-Peer-based file-sharing protocol that is used to distribute large data volumes, because it reduces the server load. When a file is downloaded using BitTorrent technology, the file is not transmitted as a single unit but in pieces of data sourced from all devices connected to the network.

Blockchain technology, which was developed primarily for alternative currencies, stands out from other technologies due to its unique data structure, which – due to the transparency it offers and its decentralised design – is particularly tamper proof. The data is stored at many separate locations and regularly compared. Blockchain technology enables (pseudo)anonymous transactions and communication – a feature that also makes the technology attractive for malign actors.

Conspiracy theories are attempts to explain a phenomenon by invoking a sinister plot orchestrated by powerful actors. Conspiracies are painted as secret or esoteric, with adherents to a theory seeing themselves as the initiated few who have access to hidden knowledge. Supporters of conspiracy theories usually see themselves as in direct opposition to the powers who are orchestrating the plot which are typically governments or figures of authority.

Deplatforming refers to the blocking of social media accounts and groups. It regularly results in these groups losing audience reach and revenue sources for their agendas. At the same time, deplatforming and the fear of accounts and websites being blocked or deleted has contributed to the emergence of alternative online platforms.

Disinformation is defined as false or misleading content that is spread with the intent to deceive, or secure economic and/or political gain, and which may cause public harm. When referring to such content that is spread unintentionally, we will be using the term **misinformation**.

Extremism is the advocacy of a system of belief that claims the superiority and dominance of one identity-based 'in-group' over all 'out-groups.' It advances a dehumanising, 'othering' mindset incompatible with pluralism and universal human rights. Extremist groups pursue and advocate a systemic political and societal change that reflects their world view. They may do this through non-violent and more subtle means, as well as through violent or explicit means. Extremism can be advocated by state and non-state actors alike.

The **Fediverse** is an attempt to create a decentralised alternative to major social networks. The Fediverse includes micro-blogging, video and image-sharing services. The different servers within the Fediverse can communicate with one another, provided the services are using the same network protocol.

Foundation models are a recent development, in which AI models are developed from algorithms designed to optimise for generality and versatility of output. These models are often trained on a broad range of data sources and large amounts of data to accomplish a wide range of downstream tasks, including some for which they were not specifically developed and trained. The foundation model can be unimodal or multimodal, trained through various methods such as supervised learning or reinforced learning.

Harmful content and behaviours refer to a broad spectrum of online activities that can have a negative impact on human rights, society and/or democracy. These can include targeted harassment of individuals, incitement of violence against a particular group or the spreading of disinformation and harmful conspiracy theories. In some instances, the risk of harm may be intrinsic to the content itself, with the risks exacerbated by amplification; in others, the harm may be caused by aggregate patterns of behaviour rather than the nature of the content itself. Depending on the geographic and legal context, different forms of harmful content and behaviours may or may not be illegal. Depending on the platform, these also may or may not be covered by a company's 'Community Guidelines', standards or rules.

Hate is understood to relate to beliefs or practices that attack, malign, delegitimise or exclude an entire class of people based on protected or immutable characteristics, including their ethnicity, religion, gender, sexual orientation, or disability. Hate actors are understood to be individuals, groups or communities which actively and overtly engage in the above activity, as well as those who implicitly attack classes of people through, for example, the use of conspiracy theories and disinformation. Hateful activity is understood to be antithetical to pluralism and the universal application of Human Rights.

Information manipulation describes a mostly non-illegal pattern of behaviour that threatens or has the potential to negatively impact values, procedures, and political processes. Such activity is manipulative in character and conducted in an intentional and coordinated manner.

Instance refers in this context to an online platform that was set up using PeerTube or other Fediverse software. You can create accounts and upload content on instances, just like on conventional online platforms. Each instance is managed independently but can communicate with other instances via optional networking functions.

Non-fungible tokens (NFTs) are units of data stored on a digital ledger that keeps records of the purchase and prevents forgery. They are unique assets in a digital world, meaning that they can be sold and bought like tangible property, and can be seen as virtual proof of ownership.

Radioactive data refer to marks (data isotopes) that remain through the learning process and that are detectable with high confidence in a neural network.

Social bots work from accounts on online platforms. They are computer programmes that, once activated, operate automatically without human input and are used, for example, to share, like or comment on posts.

Virtual Reality (VR) is a technology that provides almost real and/or believable experiences in a synthetic or virtual way, while **Augmented Reality (AR)** enhances the real world by superimposing computer-generated information on top of it. A **Mixed Reality (MR)** experience is one that seamlessly blends the user's real-world environment and digitally created content, where both environments can coexist and interact with each other. **Extended reality (XR)** is a collective term used to encompass technologies like VR, AR and MR.

Executive Summary

This policy paper provides an overview of relevant findings on the risks of harm of emerging platforms and technologies and identifies a series of policy implications. The paper analyses 'decentralised', 'generative', and 'immersive' platforms and technologies regarding their impact on disinformation, hate, and extremism. However, it should be noted that the described trends can be interdependent and that the convergence of emerging platforms and technologies is advancing rapidly. With this, their respective risks may also converge. This raises important questions about the manifestation of future harms and the potential for more visceral and extreme impacts. Consequently, policymakers and regulators must recognise the specific risks of harm and support targeted initiatives to mitigate these risks according to their nature. In addition, they need to consider converging risks in the development and enforcement of new and already existing policies. For this, transnational and cross-functional cooperation between policymakers and regulators will be crucial.

Overall threat landscape

- The decentralised social web has become a conducive environment for far right and conspiratorial milieus to share harmful content and engage in harmful behaviours. PeerTube instances, for example, allow for full individual control over content moderation and the distribution of harmful content via different interconnected services, while at the same time ousting other users with different ideologies from the information space. Similarly, on Odysee, malign actors can support their harmful content and behaviours through new monetisation options.
- Large language models (LLMs) can generate a wide variety of unpredictable outcomes, and currently offer significant potential for exploitation by malign actors. Researchers have observed that training data corpora can contain biases or stereotypes, and that even simple prompts can generate harmful content, including misinformation. Technically advanced actors may also exploit the code-generation functionality of LLMs for information manipulation. Unmediated access to consumerfacing end applications can further contribute to deceiving end users.
- Extended Reality (XR) may enable more severe versions of existing harmful content and behaviours. For example, XR could enable an unprecedented level of emotional manipulation and more convincing propaganda (e.g., through avatars), as well as make harms more physical. In addition, XR may become a retreat for malign actors to recruit, finance, or plan operations. As such, XR also offers a range of opportunities for the integration of the decentralised social web and LLMs, including their specific risks of harm, thus further increasing its own potential impacts.

Recommendations

- Policymakers and regulators must clarify which existing regulatory regimes apply to decentralised social web services, and which approaches to enforcement are applicable. Requirements for service providers to appoint in-country representatives can be considered an important policy element to achieve initial accountability. Policy enforcement must rely on improved international coordination and public pressure, as well as regulators proactively supporting provider compliance (e.g., through the development of compliance plugins).
- Risks of harm from LLMs can be experienced in a variety of consumer-facing applications. To address this, policymakers and regulators must define new rules for access, accountability, liability, safety, and detection of LLMs. Self-regulation can provide an interim approach until new rules come into force. At the same time, regulators must be aware of the tactics and techniques used by malign actors to exploit LLMs for their strategies (e.g., for information manipulation).
- Policymakers and regulators must define what constitutes risks of harm in XR environments and ensure there is an applicable regulatory or co-regulatory framework in place. To this end, they must review existing platform and technology regulations (e.g., the EU's Digital Services Act and AI Act, Australia's Online Safety Act 2021), and national criminal codes and potentially develop a XR specific harms and crime taxonomies. In addition, standards for evidence gathering, reporting, and moderation must be developed in a multistakeholder dialogue.

Introduction

On 11 April 2022, China's Supreme People's Court (SPC) published a case in which the provider of an AI application used the image of a natural person without permission and created a virtual character as an 'AI companion', identified by the person's name and portrait, and enabled the creation of interactive content.[1] The case illustrates the ongoing trend of increasing convergence of digital technologies with consequences for our human rights such as personality rights. The information space has always been characterised by technological innovation and inter-relations between media and society,[2] leading to new benefits but also risks of harm. Along with the perception and impact of media, governance approaches have also progressed. As a result, the legal framework was tightened in many contexts (e.g., the Harmful Digital Communications Act (2015) in New Zealand, and the Network Enforcement Act (2017) in Germany). It is not yet fully clear to what extent these new regulatory frameworks have impacted fundamental rights. However, emerging platforms and technologies like the decentralised social web, large language models (LLMs), and extended reality (XR) are increasingly exploited by malign actors, including the online extremist and propaganda ecosystems, that disseminate harmful content and/or engage in harmful behaviours. This is also recognised by the Global Coalition for Digital Safety, which acknowledges that the technologies and platforms mentioned "may give rise to new forms of harm or exacerbate existing ones".[3] In line with this evolving threat landscape, policymakers and regulators around the world will need to analyse to what extent their approach to risks of harm is future-proof. Building on the discussions of a Digital Policy Lab (DPL)[4] working group on 'Emerging Platforms and Technologies' between May and June 2023, this policy paper will discuss the risks of harm and policy implications of emerging 'decentralised', 'generative' and 'immersive' platforms and technologies. This analysis aims to provide an overview of possible negative consequences of these, as well as impulses for policy development, based on current research findings and examples. The analysis is primarily limited to selected risks of harm related to disinformation, hate, and extremism. In three sections, the paper will explore the three types of emerging platforms and technologies: section one focusing on the decentralised social web, section two on LLMs, and section three on XR. Each section includes explaining boxes on key terms.

Decentralised, generative and immersive: the evolving online extremist ecosystem

Section 1: Decentralised social web

The social media landscape has seen a rapid transformation over recent years. As the dominant platforms like Facebook, X (formerly Twitter), and YouTube have increasingly come under public scrutiny for their role in facilitating the spread of disinformation and hate, alternative digital services have been emerging. Some of these new services—like the video platform Odysee—have sought to provide a safe refuge for conspiracy ideologues, far-right extremists, and any other malign actors that feel their content is unfairly censored on major online platforms. Others have emerged to provide an alternative to the surveillance capitalism model of 'Big Tech', aspiring to create a social media universe free of corporate interests – the so-called Fediverse. Both Web3 services like Odysee and the Fediverse are often described as 'decentralised' alternatives to the dominant large online platforms. Yet, both differ in terms of technology and wider vision, while interfaces between their information spaces are emerging.[5] The following section will explore the potential exploitation of the decentralised social web related to disinformation, hate, and extremism, before assessing respective policy implications.

Explainer: Web3

Odysee is part of the Web3 movement, which journalist Gilead Edelman described as "a decentralised online ecosystem based on the blockchain." He continued that "[p]latforms and apps built on Web3 won't be owned by a central gatekeeper, but rather by users, who will earn their ownership stake by helping to develop and maintain those services".[6] At the heart of Web3 is blockchain technology, which users can use to authenticate themselves and their content, as well as cryptocurrency, through which participants in Web3 (infrastructure providers as well as regular users) can receive financial rewards. The emphasis on financial transactions free from government interference has made the Web3 popular among libertarian groups. Odysee is a video hosting platform that has been marketed by its developer LBRY as a YouTube alternative, which enables streaming and file downloads. Odysee is built on the LBRY protocol, a decentralised filesharing network that incorporates blockchain and BitTorrent technologies and uses LBRY Credit (LBC) as currency. As a result of a New Hampshire court decision in July 2023, LBRY had to announce its closure, thus calling into question the future of Odysee and comparable Web3 services.[7]

Explainer: The Fediverse

The Fediverse shares the same goal of a decentralised social web based on peer-to-peer networks, albeit without the emphasis on financial rewards or blockchain technology. Instead, the Fediverse evolved out of the so-called free software movement. Like proponents of open-source software, free software allows anyone to freely use, change, and distribute software. Yet, free software proponents also stress ethical principles like avoiding proprietary software that hinders cooperation. In the Fediverse, protocols like ActivityPub allow different platforms (also called instances) to speak to each other, hence emphasising interoperability over competition. This means that users can interact with each other and their content, even if they are not on the same instance. In this way, users on Mastodon (an X-like microblogging service in the Fediverse) can comment on a video on PeerTube (a Fediverse software that enables the creation of YouTube-like video services) without having to create a new account. Notably, the Fediverse is already being used by the official bodies of the European Union[8] and the German government[9], and is endorsed by their respective data protection authorities.

The threat landscape

While there has been a strong research focus on the risks of harm from large online platforms and their use in the past, less attention has been paid to the decentralised social web. According to Tech Against Terrorism, the exploitation of decentralised services by terrorist and violent extremist (TVE) actors is still primarily experimental, with these services being used alongside or as backups to conventional, centralised platforms and services.[10] However, as the number of users has increased, interest in the potential risks of harm of Web3 services and the Fediverse has grown, which is reflected in new initiatives to access and study alternative information spaces. For example, in September 2022, New Zealand, the USA, X, and Microsoft, announced an investment in a technology innovation initiative in partnership with OpenMined under the banner of the Christchurch Call for Action.[11] The goal of the initiative is to support the independent study of the impacts of algorithms and their interactions with users, including across multiple platforms and types of platforms.

Platform migration and echo chambers: Both Web3 services like Odysee and Fediverse services such as PeerTube have seen a growing user base over the past years. This specifically applies to far-right and conspiracy milieus, for whom these services can provide a conducive environment to spread harmful content, including disinformation and hate, and engage in harmful behaviours.[12] Actors and movements 'deplatformed' from YouTube, like the German 'Querdenken' movement or known far-right agitators, have quickly found a new home in the decentralised social web. For example, by using PeerTube, far-right extremists can create instances that they alone can control. The content disseminated there can then only be removed by taking the entire servers offline. To achieve ideological dominance, there is evidence of users who do not comply with their ideals being bullied.[13] In the wake of what can be described as 'community capture', these information spaces are increasingly becoming ideologically uniform 'echo chambers'. However, the separation of these alternative information spaces makes them harder to regulate, enabling malign actors to exchange ideas without interruption or challenge.

Cross-sharing harmful content: Users on Fediverse instances can share content with users of other instances. However, unlike with very large online platforms such as Facebook or X, on the Fediverse there is no central controller who can moderate the flow of information across instances or delete unwanted content.[14] The root of this challenge is the Fediverse's key benefit and feature, namely, its decentralised model – because "there is no centralised Fediverse authority, there is no way to fully exclude even the most harmful content from the network".[15] As one commentator explains, reliably deleting content from a decentralised network "is just not possible".[16] While this risk of harm is not absent from very large online platforms given that other users may have copied the post before deletion, Mastodon, for example, makes a copy for each user that views the post. On such Fediverse platforms, the 'right to be forgotten' under Article 17 of the EU's General Data Protection Regulation (GDPR) may be practically unenforceable.

Monetising harmful content: Odysee allows users to monetise their content and use of the platform through different reward, bonus and boost functions. For example, users can earn LBCs by creating a channel, uploading videos, following other profiles, or gaining followers. The earnings depend on the achievement of different levels (e.g., 'Master of Views'). The possibility of advancing to certain levels and the design of the platform have elements of a gamification strategy, according to which the motivation of users to use the service can be increased through game-like mechanisms.[17] This is particularly attractive for those actors that have been demonetised or deplatformed from larger online platforms such as YouTube. An ISD study based on 53 German-language Odysee users from the far-right and conspiracy milieus found that accounts had received 1,652,786.96 LBRY Credits in total (corresponding to 122,306 USD according to the average closing price for LBC from January 2022 to May 2023) since the wallets were set up.[18] The volatile nature of cryptocurrencies might have resulted in the earnings being greatly diminished since the time of analysis; however, the numbers show that incentivised platforms like Odysee can create additional income for malign actors. It was further found that profitable far-right extremist profiles disseminated videos promoting

conspiracy theories such as 'The Great Reset' and 'QAnon'. Investigated videos also contained antisemitic statements, historical revisionism, and climate change denial. In the researched video sample, core extreme-right topics, such as Holocaust Denial, received less support than videos discussing current political issues.

Policy implications

The regulation of alternative information spaces has been on the minds of policymakers and regulators for some time already. This is due to the phenomenon of platform migration and echo chambers described above, which are important in understanding the way extremist movements develop and operate online. While the decentralised social web may help achieve certain policy goals like open source first, compliance with data protection or interoperability, it also means a largely uncontrolled information space that is open to exploitation by malign actors. Although policies often already exist, there are still considerable challenges with enforcement, as some Web3 and Fediverse service providers deliberately try to evade regulatory scrutiny.

Considering new monetisation: While some established large online platforms already remunerate user activity (e.g., Super Chat feature on YouTube), gamified experiences on Web3 platforms like Odysee pose a new level of risk of harm, regarding the profitability of disinformation and hate through gamified experiences.[19] The technology and financial incentive structure of these types of platforms must be considered when crafting new policies or updating existing frameworks. The latter may not only include those frameworks regulating platforms and technologies but also counter-extremism initiatives. At the same time, existing policies may already offer possibilities to counter the monetisation of content and behaviours by malign actors on Web3 services. These provisions would need to be enforced effectively, as seen in the case brought by the U.S. Securities and Exchange Commission (SEC) against LBRY.[20]

Recognising technological bluffs: The descriptor 'decentralised' is often used as a marketing tool and may be used as an excuse not to comply with relevant legislation. However, despite claims of being decentralised, the video platform Odysee can, and does moderate content.[21] Though using blockchain technology, Odysee can still delist channels or geo-block content and therefore make content virtually inaccessible for most users. The content may still be accessible for tech-savvy users via the blockchain, but it can no longer be viewed on Odysee. This means the platform is, in principle, able to comply with regulations like duties to establish notice-and-takedown mechanisms.

Determining types of services: Many PeerTube instances allow the sharing of user-generated content. However, some instances also offer editorial formats. Others in turn only allow certain individuals or media organisations to upload content. These different types of use mean that some instances could legally be considered social networks or online platforms, while others are considered publishers of editorial content or media platforms. These various categories may have different legal obligations in different jurisdictions. Moreover, some regulatory frameworks require the services to make a profit (e.g., Germany's Network Enforcement Act (2017)). However, for most PeerTube instances it is unclear what the business relationship is between the operators, the content providers, and the users. Moreover, the corporate structure behind the instances is often unclear. There is clarity, however, when it comes to funding, which usually takes place via donations and often in the form of cryptocurrencies.[22]

Improving enforcement: The decentralised nature of Fediverse services means there is no central platform authority that can be appealed to for taking down content. Instead, each server administrator moderates content locally. Some instances let users view potentially illegal content from linked instances, even if that instance itself does not host such content. This raises questions about enforcing legal notice-and-takedown procedures for illegal content within decentralised networks. Moreover, some instances appear to be run by individuals or organisations in jurisdictions without noteworthy policies for the digital information space, sometimes seemingly using shell companies. Following that, policies requiring providers to establish domestic legal representatives for the delivery of legal documents should be considered by policymakers. However, in 2021, the German Federal Government ran into problems when trying to deliver notices to Telegram, despite a corresponding duty in the

NetzDG. After the unsuccessful delivery (Telegram, which is officially headquartered in Dubai, did not respond), the notices could finally be 'delivered' through publication in Germany's Federal Gazette. Together with the pressure exerted via the media and international coordination, Telegram ultimately appealed, which is why a court case is underway in 2023.[23]

Building on existing approaches: Framasoft, the French non-profit developing PeerTube, uses indexing to actively remove problematic instances from their search functionality if they maintain illegal content under French law.[24] If indexing standards are tightened, this could be an adequate approach against the dissemination of illegal content, as PeerTube instances can only find a wider audience once they have been included in searchable indexes. In this way, extremist content of problematic instances can thus be made less accessible. Additionally, community-run block lists allow server hosts to avoid federating with problematic instances and therefore avoid spreading extremist content accidentally by allowing them to participate in the federation system. In this way, extremist instances may still be accessible directly, but their content will no longer be visible on the wider network.

Driving capacity-building: Policymakers and regulators could support the development of open-source plugins that Fediverse instance owners could use to streamline content reporting and moderation processes compliant to various regulations, such as the EU's Digital Services Act (DSA) or Australia's Online Safety Act. This will not solve the problem of content moderation, especially for instance owners that have no additional personnel, but it would show willingness on both sides to comply with transparency requirements as far as possible and initiate a dialogue between the Fediverse community and regulators. Policymakers and regulators may also encourage and support civil society to offer training to instance owners on best content moderation practices and standard development.

Section 2: Large language models

In an open letter published in March 2023, the Future of Life Institute (FLI) called on AI companies to pause the training of AI applications more powerful than GPT-4.0 because of "profound risks to society and humanity"[25]. While the letter found support from some researchers, others criticised it for prioritising imagined apocalyptic scenarios over the risks of harm posed by the widespread use of Generative AI already.[26] However, currently accessible AI systems, which are based on large language models (LLMs), and their use entail both opportunities and negative consequences for individuals and society. The following section will highlight select risks of harm inherent to these models and the potential for their exploitation by malign actors with a focus on disinformation, hate, and extremism. Several related policy implications will then be discussed afterwards.

Explainer: Generative AI

Whereas artificial intelligence (AI) is an umbrella term attributed to systems and technologies that mimic human intelligence, Generative AI refers to AI applications that can generate new code, text, images, audio, video, and multimodal simulations in response to prompts, as well as underlying language models (LMs) on which applications can be built. OpenAI's ChatGPT, for example, generates text. DALL-E creates realistic images and art. ChatGPT and DALL-E are based on GPT-4.0, 3.5 or 3.0, and DALL-E (a version of GPT-3.0) respectively, which are examples of large language models (LLMs). LMs are architectures of neural networks – series of algorithms that, loosely speaking, mimic the operations of an animal brain, and can recognise complex patterns. These neural networks are comprised of node layers, "containing an input layer, one or more hidden layers, and an output layer."[27] Neural networks are particularly useful for clustering and classifying information. The more node layers, the more capable the neural network is of handling very large and complicated datasets and discovering patterns within unlabelled and unstructured data. Many of such LLMs can be reused in countless downstream AI applications.

The threat landscape

The full impact of LLMs on individuals and society is not foreseeable. Yet, researchers[28], policymakers[29], law enforcement[30] and industry[31] are increasingly concerned with the risks of harm they can pose. For example, Weidinger et al. derived 21 risks from LMs across six potential risk areas, including 'discrimination, exclusion and toxicity', 'misinformation harms' and 'malicious uses'.[32] Others focused specifically on audio, text, visual, or multimodal content that has been generated or modified by AI. The Partnership on AI, for example, created a list of risks of synthetic media and responsible practices, stating that, "as synthetic media technology becomes more accessible and sophisticated, its potential impact also increases"[33]. It is therefore essential that mitigation measures address both the supply and demand side, while more research is still needed to fully understand the specific risks of harm of AI-powered information manipulation.[34] Providers constantly seek to improve their applications' accuracy. However, the evolving capabilities of LLMs remain uncertain, as there are currently no reliable techniques for "steering the behaviour of LLMs".[35]

Inherent biases: Toxicity, including hate, was identified as a prevalent issue in both large LLMs and web text corpora.[36] These models can sometimes assign high probabilities to utterances that constitute harmful content such as misinformation or hate. For example, if the training data does not respect minority views, there is also a risk that LLM distributions will reinforce majority over minority views and values. Researchers showed how LLMs displayed undesirable stereotypes such as persistent associations between Muslims and violence.[37] In a more recent large-scale analysis of half a million generations from GPT-3.5, researchers also found that LLMs can be significantly toxic when assigned personas such as 'a bad person'.[38] Concerning misinformation, even advanced LLMs do not reliably predict true information – these models emit detailed and correct information in some circumstances but then provide incorrect information in others.[39] These findings not only show that LLM outcomes can be inherently toxic or misleading (depending on the context) but also that seemingly innocuous prompts by everyday users can generate such content. In this context, especially AI applications that are based on the corpora of platforms whose content has already been identified as toxic or misleading should be assessed critically.

Intentionally malicious use: LLMs are vulnerable to intentionally malign uses to generate harmful content or perform harmful behaviours. In an experiment, GPT-3.5 was directed to respond to a series of leading prompts relating to 100 verifiably false narratives.[40] The chatbot generated 80 of the 100 false narratives. The developer subsequently promised improvements. However, GPT-4.0 was found to advance prominent false narratives even more frequently and persuasively than GPT-3.5.[41] Aside from the potential exploitation of AI-generated text by malign actors, compellingly realistic AI-generated deepfakes can enhance the effectiveness of information manipulation, exacerbating familiar threats to the quality of public discourse and the digital information space.[42] Images and video content is still frequently used in foreign information manipulation incidents.[43] This is because such content is appealing as well as cheap and easy to produce. Consequently, malign actors will look closely at the cost difference between human-generated and AI-generated content.[44]

Generation of code: LLMs can be utilised to analyse, debug and generate computer code. Consequently, there is speculation about the potential for LLMs to assist malign actors to generate code more efficiently for nefarious purposes including cybercrime, the operation of social bots, or politically motivated cyber-attacks. Politically motivated cyber-attacks are, of course, nothing new: In November 2022, the European Parliament website was targeted by a 'sophisticated' cyber-attack 'moments after' a vote declaring Russia to be a sponsor of terrorism.[45] While there may be limits as to how malign actors with limited resources and technical knowhow harness the evolving capabilities of LLMs, more advanced and well-resourced actors, such as foreign intelligence agencies, could exploit the models in combination with other technologies to enhance, automate, and scale sophisticated information manipulation campaigns.[46] At the same time, LLMs could also be used to enhance cyber security by examining code for vulnerabilities and debugging code.[47]

Automated distribution: LLMs could be exploited in combination with techniques that enable automated distribution of content. For example, social bots have already been integrated within information manipulation campaigns and used to amplify conspiracy narratives, for example.[48] In March 2023, investigative journalists published leaked emails and other documents ('Vulkan

Files') that evidence the development of sabotage software by the Russian company NTC Vulkan. Commissioned by the Russian Federal Government, some developed programmes were already able to automatically generate and disseminate content.[49] The use LLMs to "re-produce language patterns" and thus convincingly impersonating target individuals and groups could therefore make such programmes even more sophisticated.[50] Consequently, combining techniques for automated distribution with LLMs will open the door to tailored and personalised information manipulation, including disinformation, at scale.[51] Moreover, LLMs may undermine current techniques used to detect information manipulation by reducing or altogether removing reliance on copy-and-pasted text.[52]

Improved persuasiveness: Misinformation and information manipulation have always existed, especially in politics. However, their impact on individuals and society ultimately depends on their ability to persuade listeners or readers to believe a particular message. In a game-design experiment, it was demonstrated that LLMs can hold sophisticated conversations and dialogues with humans in a highly convincing way.[53] In another experimental study, it was found that messages generated by GPT-3.0 were even persuasive across several policy issues, including an assault weapon ban or a carbon tax.[54] A study from June 2023 involving 697 participants confirmed these results.[55] GPT-3.0 could not only produce accurate tweets that were easier to understand, but it could also generate more persuasive synthetic misinformation. The authors also showed that respondents were not able to distinguish between tweets generated by GPT-3.0 and those written by real users.

Unmediated access: The persuasiveness of AI-generated content may also be enhanced by the fact that access to LLM outputs is often unmediated, as there are no comparative results as with conventional search engines like Google or Bing. This is especially problematic when LLMs provide correct information for most search requests, which may lead users to overly trust the predictions, including misinformation in single cases.[56] Given the potential persuasiveness of LLM outputs, it is not surprising that people are particularly fearful of AI-generated misinformation.[57] However, the impacts of LLMs on media trust have yet to be demonstrated and further studied.[58]

Policy implications

Policymakers and regulators seeking to regulate LLMs are faced with novel and urgent questions that are different to those posed by architectures of neural networks with narrow and intended use cases or other kinds of software. This in part relates to the unpredictability of such models, which may contain or evolve capabilities that are dangerous and accessible to malign actors.[59] While new efforts to regulate AI and especially Generative AI are gaining momentum in mid-2023 among parliamentarians in the US[60], seven of the most influential AI companies, including Amazon, Google, Meta, Microsoft and OpenAI, agreed with the White House in July 2023 to immediate voluntary commitments to manage the risks of harm posed by AI[61]. These commitments underscore the principles of 'safety', 'security', and 'trust', but apply only to LLMs more powerful than GPT-4.0, Claude 2, PaLM 2, Titan and, in the case of AI-generated images, DALL-E 2. However, as discussed in this paper, the current generation of AI applications already presents severe risks of harm for which mitigation measures are urgently needed.

In June 2023, the European Parliament (EP) has already adopted its negotiating position on the AI Act, including new rules for foundation models – a term used parallel to LLMs.[62] In general, the currently proposed AI Act is likely to be underpinned by a 'risk-based' approach, classifying systems according to their potential to infringe on public safety and fundamental rights. The rules proposed by the EP include additional transparency requirements for foundation models, like disclosing that the content was generated by them and designing the models to prevent it from generating illegal content. However, as of August 2023, the final text of the AI Act may yet change and will not enter into force until 2025. This is why the European Commission is trying to persuade AI developers to pre-empt the AI Act by entering a voluntary AI Pact.[63] At the same time, policymakers and regulators around the world are considering which existing laws might already apply.

Determining accountability and liability: Establishing individual responsibility for harmful outputs and shortcomings of AI is difficult as its systems are opaque, unpredictable, and involve a multitude of actors and resources.[64] Accountability, therefore, is the 'cornerstone' of AI governance.[65] These complications

also underpin uncertainty around whether providers of LLMs are or should be legally liable for their outputs. While providers of online platforms have limited liability for user-generated content in most liberal democratic countries (e.g., Section 230 of the US Communications Decency Act; Art. 4 DSA), this may not apply to providers of LLMs, as LLMs are likely to be assessed (at least partially) as content creating services rather than intermediary services.[66] However, it is challenging for claimants to prove whether certain harm has occurred due to misconduct on the part of the provider, as advanced LLMs lack explainability (e.g., the question of how certain outputs were generated). Therefore, policymakers should consider policies that require providers to increase model transparency (through, for example, regular transparency reporting, model or system cards[67]) and find alternative ways to define the misconduct of providers in certain cases (e.g., infringement of due diligence duties).

Limiting access: The cost difference between human-generated and AI-generated content is particularly relevant to malign actors. A central factor here is the extent to which malign actors can integrate LLMs in their own applications through developer Application Programming Interfaces (APIs). Policymakers and regulators could oblige providers to introduce robust vetting criteria and perform background checks on the potential use of their services by malign actors. However, it is still unclear whether vetting procedures could eventually prevent malign actors from accessing developer APIs, and if so, whether they can simply move to open-source models.[68] Moreover, it is also possible for governments or other well-resourced actors to train and operate their own systems.[69] This would enable such actors to bypass guardrails that, for example, established LLMs have in place around filtering harmful content.

Inclusive design: As shown in this policy paper, biased training corpora can lead to toxic content, including hate, and misinformation. Providers of LLMs should therefore be obliged to put in place proportionate and effective measures to counter inherent toxicity and misinformation in LLMs. This may include, for example, changing the training data distribution, training a classifier and using it to reduce the probability of harmful content, or sensitise human supervisors to harmful content Providers of LLMs and products built on them could further be obliged to publish public-facing specification reports which include toxicity and misinformation stress tests to inform users.[70]

Safety by Design: Safety by Design is built on three core principles: service provider responsibility, user empowerment and autonomy, transparency and accountability. Policymakers and regulators should commit providers of LLMs and respective AI applications to uphold these principles by ensuring they incorporate safety measures at every product lifecycle stage. This must involve consulting stakeholders from multiple sectors and collaborating with the user community, including those who are typically under-represented or who may be at greater risk of online ha.

Protecting rights: AI-generated deepfakes can enhance the effectiveness of information manipulation. However, AI-generated content may in certain circumstances directly infringe fundamental rights. In 2008, a footballer successfully sued Electronic Arts in a court in Hamburg, Germany, preventing the FIFA videogame from using his likeness without consent.[71] Crucially, the harm derived from the claimant's "right to choose how his name might be used" rather than from commercial considerations.[72] Enforcing fundamental rights may thus be another option to prevent malign actors from exploiting LLMs. Yet for this to be achieved, questions of liability need to be clarified first.

Improving detection: Labelling AI-generated content through, for example, digital watermarking[73], is a frequently discussed topic among policymakers and regulators.[74] In practice, however, identifying AI-generated content is a challenging task.[75] On the one hand, building LLMs with more detectable outputs is technically difficult (e.g., directly manipulating LLM parameters to create statistical fingerprints; or training models with 'radioactive data') and requires further research and coordination among developers. On the other hand, spreading 'radioactive' data directly on the internet, where it would likely be tapped by those wishing to train LLMs, raises ethical concerns as large amounts of data would need to be disseminated. Consequently, providers of LLMs, online platforms, policymakers, regulators, and researchers must further align to improve the detection of AI-generated content. The Coalition for Content Provenance and Authenticity (C2PA)[76] and the Partnership on AI[77] are already promising initiatives in this regard.

Section 3: Extended reality

The Pew Research Center consulted more than 600 experts between February and March 2022 to hear their predictions about the trajectory and impact of the Metaverse by 2040. A notable share of them argued that the embrace of Extended Reality (XR) in people's daily lives will be centred around Augmented Reality (AR) and Mixed Reality (MR), not in a more-fully-immersive Virtual Reality (VR). They warned that these technologies can "dramatically magnify every human trait and tendency – both the bad and the good."[78] For example, Toby Shulruff, a senior technology safety specialist at the National Network to End Domestic Violence, noted that "like other technologies, XR does not solve human problems like bias, fear or violence" but it instead "accelerates and amplifies what is already present in society". He further warned that there was a real possibility that those who were 'plugged in' would become increasingly untethered from the world around them.[79] While the extent to which people's everyday lives will be shaped by full-immersion XR remains ambiguous, we are already aware of the various challenges posed by current digital information spaces. As with online platforms, the XR-based spaces like the Metaverse will face issues concerning liability, law enforcement, content moderation, data privacy, transparency, and the protection of users' fundamental rights. This section will explore the risks of harm of the Metaverse and its use before outlining policy implications.

Explainer: The Metaverse

Coinciding with Facebook's rebranding as Meta in October 2021,[80] CEO Mark Zuckerberg announced his intention to build "the successor to the mobile internet"[81] calling it the 'Metaverse'. The Metaverse has been described as "a convergence of physical, augmented, and virtual reality in a shared online space."[82] Like other XR worlds, it thereby intends to go beyond game-like goals and gamification. There are many descriptions and analogies of the Metaverse, so it is worth breaking down what it is supposed to be. In an essay[83], released in January 2020, Matthew Ball outlined seven core attributes of the Metaverse: (1) persistent – continuing indefinitely; (2) synchronous and live – a living experience that exists consistently for everyone and in real-time; (3) an individual sense of 'presence' – the feeling of interacting with other users (for example, in an event) and objects as if they were physical with them; (4) a full-scale economy – individuals and businesses will be able to create, own, invest, sell, and be rewarded for work; (5) a merge of virtual and physical worlds – spanning private and public networks/experiences, and open and closed platforms; (6) fully interoperable – users will be able to take their avatars and digital items/assets from one platform in the Metaverse to another; and (7) populated by content and experiences – created and operated by a range of contributors. In December 2021, Meta opened access to its multiplayer VR platform, Horizon Worlds.[84] However, the Metaverse as described above does not exist (yet) and depends on several conditions including the availability of hardware and thereby means of entry (e.g., helmets, lenses, sensory suits, or even neural links), a strong internet connection, and technical standards for software to enable interoperability[85].

The threat landscape

In 2022, the Europol Innovation Lab[86] and the EU Counter-Terrorism Coordinator[87] published papers on the potential for adverse use of, crime, and radicalisation in the Metaverse. The documents review the modalities of the Metaverse, including the immersive nature, the capture of emotions, and the use of avatars, and their implications for the proliferation of harmful and illegal content, harassment and abuse as well as terrorism, especially for recruitment, financing, and training. In February 2023, the WeProtect Global Alliance[88] also published an analysis, providing an overview of the latest trends on XR and its potential impact on child sexual exploitation and abuse online. In May 2023, Standards Australia published a whitepaper, outlining key definitions related to the Metaverse, various risks, including 'human risks' and 'societal risks', and existing standards work. Soon after, in July 2023, the World Economic Forum (WEF) [89] addressed privacy and

safety concerns related to the Metaverse. While some developers seem keen to emphasise the benefits of the Metaverse (we use the term in the following as a reference to XR-based information spaces focused on social connections, and not the specific product from Meta by the same name), there are risks of harm that should be considered further.

Sexual harassment, abuse, and gender-based violence: The Metaverse is a gendered space influenced by the misogynistic and hypermasculine culture in gaming spaces.[90] The Europol Innovation Lab recognised an incident of a women describing how she was "virtually gang raped" within 60 seconds of joining Meta's Venues.[91] Gaming environments, like other social spaces, can reproduce systems of structural discrimination and inequality such as racism, sexism, and ableism.[92] Thereby, online gender-based violence on 'traditional' online platforms has been understood as a continuum of offline violence that involves many forms targeting women, girls, and marginalised gender identities, especially those with intersecting identity factors such as race, indigeneity, class, sexual identity, and sexual expression, or disability. The embodiment and sensation of presence afforded by the Metaverse can make harassment feel more intense, enabling violations of personal space and corporeal presence. This may also be particularly harmful for children who could potentially be sexually harassed and abused by deceiving avatars who initially act like child companions.[93] Likewise, the ephemerality of the Metaverse makes it difficult to report unwanted behaviours.[94] It is thus important to acknowledge the possibility of very real, impactful experiences of sexual harassment and abuse.

Psychological impact: The experience of violence in the Metaverse can have a real-life impact on users' mental health, especially for young users. Researchers found that exposure to violence at high levels or across multiple contexts in early adolescence has been linked with emotional desensitisation, which contributes to serious violence in late adolescence.[95] Desensitisation to violence is a form of habituation, a type of non-associative learning that results in a diminished response to a stimulus after repeated exposure, extending across contexts and settings.[96] For example, witnessing a fight may produce desensitisation to other types of violence in the same context, as well as violence observed in other settings (e.g., home or school).[97] Witnessing for example mass killings of avatars by terrorists in the Metaverse could produce similar desensitisation and other psychological harm.

Ideological enabler: The Metaverse could offer a space for malign actors to spread propaganda, recruit users, and exercise control over a radicalised community through events and regular meetings. Firstly, the emotional investment and ambiguous distinction between real and virtual worlds could make users more susceptible to emotional manipulation.[98] Avatars could be misused to provoke emotions and spread extremist ideology, for example, by featuring deceased terrorist leaders "in a virtual resurrection."[99] Secondly, the use of hardware that captures body-based data like eye tracking or other body motions could accelerate the means of recording biometric data, creating new ways for impersonation or selecting and targeting vulnerable user and tailoring messages to their biases That will enable malign actors to more effectively target their propaganda and recruit people. Thirdly, the Metaverse could be used for reproducing emotive historical events or creating a vision of the world (e.g., a virtual Caliphate, or a white supremacist state) to galvanise supporters. The Europol Innovation Lab asserted that such spaces could become a parallel world that undermines the rule of law.[100] For example, Nazi gas chambers have already been reported in Roblox, a platform where people can create their own experiences or mini-games and share these with other users.[101]

Financial and operational enabler: Relying on blockchain technology, cryptocurrencies and NFTs, records of digital ownership stored in the blockchain are expected to play an important role in the economy of the Metaverse.[102] Cryptocurrencies could be used for money laundering, making the monitoring of transfers – especially across borders – more difficult. Funds could be raised through the sale of artefacts such as swastikas as NFTs, which are then used to customise avatars or display their affiliation to terrorist or extremist organisations. The Metaverse could also enable training environments and scenarios, for example, practising precision shootings, hostage-taking, or even reconnaissance. The emotional and immersive nature of the Metaverse would make this kind of training more realistic and absorbing. Modelling could give terrorist organisations a tool to replicate targets from the real world to practice an attack and maximise its impact.

Policy implications

The Metaverse raises many of the same policy debates as the internet given that new features – especially its immersiveness and ephemeral nature – will need to anticipate known issues related to the privacy and user safety. Furthermore, the Metaverse will create distinct risks to individuals and society that need to be addressed in a specific manner. A paper by the Council of the EU's Analysis and Research Team, published in 2022, anticipates a potential struggle between the respective roles of government, industry, and users, which could lead to different models: a regulatory framework with an emphasis on the protection of fundamental rights; an approach focused on a free, decentralised and open internet with currency and property rights; or a business-oriented and profit-driven model backed by industry which claims ownership over the Metaverse.[103] The challenge of differing outcomes and interests highlights the need for shared approaches and models early on.

Adequacy and applicability of criminal codes: Reports of sexual assault in the Metaverse pose new questions concerning the requirements for 'traditionally' physical crimes. For instance, rape requires a physical act, while an avatar is virtual. However, it should be argued that the immersiveness of the Metaverse can result in the very real (emotional) experience of rape, reiterating questions about the requirements of existing laws to quality for the legal terms of rape. It will be important to review existing criminal codes, define what constitutes criminal behaviour in the Metaverse and to have new legislation to provide the means to prosecute these offences. A legal review should draw on multiple perspectives from civil society, industry, law enforcement and science, and assess both the adequacy and applicability of existing criminal codes and legislation, especially in terms of psychological and bodily harms and the delineations between physical and virtual. Such a review could develop the conditions under which an avatar can be equated with a person and discuss whether any observed harms should be criminalised via new legislation. Policymakers, regulators and law enforcement agencies could develop a Metaverse-specific harm and crime taxonomy to proactively identify new types of harms and crimes.[104]

Investigating crimes and gathering evidence: The Europol Innovation Lab stressed that the ephemeral nature of the Metaverse could lead to a lack of virtual traces and difficulties in gathering evidence of criminal incidents.[105] Notably, the Metaverse will face challenges in recording any data of incidents given the ephemeral user experience and potentially short response times when users report incidents. Another obstacle to criminal investigations is the challenge of entering restricted 'premises' in the Metaverse that require a key or NFT. Additionally, given the difficulty of establishing the location of users (or the access device used) and thereby challenges in finding a perpetrator in the physical world, law enforcement will face challenges in terms of identifying the users' real identities as well as establishing jurisdictional authority. Law enforcement authorities will need to coordinate to consider redress capabilities and remedies across jurisdictions. Such effort should also work toward developing a privacy-orientated data supply chain.[106]

Developing moderating practices: The Europol Innovation Lab assessed that "Just patrolling in a virtual car driving around a Metaverse will probably not work very well with potentially endless worlds, both for deterrence as well as for being approachable."[107] Known challenges of moderating online content, especially in terms of balancing moderation with freedom of expression and human rights, will be amplified in the Metaverse as companies will need to shift their focus to ephemeral user behaviour rather than content. Challenges will relate to the use of human and AI-based moderation, including questions of resources, data biases and algorithmic discrimination. Companies will need to find privacy-preserving ways to moderate interactions while enabling real-time enforcement of community guidelines. Private spaces will require different forms of moderation. For example, should a person's speech in their virtual home be subject to the same moderation as in a Metaverse's public town square? Addressing the distinction between public and private spaces online (which is already a difficult challenge for policymakers)[108] will require a clearer understanding of what content and conduct should be allowed where in the Metaverse. Equally, geographies have different legislation that will require varying levels of consideration for conduct and content moderation, for example, some countries may prohibit certain gestures or alcohol consumption.[109] Users should also have the option to report behaviours that breaches community guidelines in real time.

Safety and Privacy by Design: Policymakers and regulators should review (co-)regulatory and voluntary frameworks that cover online platforms, especially in terms of the provisions obligating service providers to assess and mitigate risks of their services to assess their applicability to the Metaverse. In the EU, this could include the DSA and the AI Act (forthcoming 2025). Additionally, providers of XR services should consider standardising safety tools like muting, blocking, and other safety resources in Metaverse contexts. Australia's eSafety's global Safety by Design initiative, for example, provides three overarching principles to guide and support the industry to enhance online safety practices: service provider responsibility, user empowerment and autonomy, and transparency and accountability. Meta notes that users in the Horizon Worlds could use a 'safe zone', which it describes as "a personal space where you can take a moment away from other people and your surroundings." By selecting the 'shield' icon, users can report the world they are currently in; view, report or block nearby users; mute or unmute someone; or go to their "personal space.[110] Meta further introduced a 'personal boundary' feature, a roughly four-foot distance between your avatar and non-friends.[111] It includes restricted settings when two users meet for the first time, for example, if one user's 'personal boundary' is off while the other's is set to on for everyone, then the Metaverse will establish a four-foot space between both. While these changes may increase users' feeling of safety, they only came after reports of sexual assault. This reiterates the need for the industry to proactively assess risks and create mitigation measures for addressing how the online harms of today's internet might be exacerbated at scale. Ultimately, the onus should not be on the users, but on the provider to protect their users.

Promoting standard-setting: The Metaverse comprises a diverse landscape of different applications owned, developed and operated by individual developers and made available to users via hardware-specific stores. As a result, policymakers and regulators should consider developing and applying horizontal rules to different services when regulating the Metaverse.[112] Research conducted on Reddit suggests that macro-level norms can also "help moderators of new and emerging communities shape their regulation policies during the community's formative stages"—but only if the presence of such site-wide norms is known.[113] Shared norms and standards for accepted behaviour in the Metaverse are still emerging. Some users may unknowingly violate expectations for accepted behaviours, while others aim to cause intentional harm.[114] Like-minded policymakers and regulators should actively engage providers, who are building the technologies of the Metaverse, to hold them accountable early on. The exchange should include developers behind Metaverse applications to discuss their policies associated with the use of VR hardware. At the EU level, the European Commission has called for public feedback to develop a "vision for emerging virtual worlds," based on "respect for digital rights and EU laws and values."[115] The call highlights the need for "cross-cutting enablers such as the appropriate governance models to ensure EU leadership in virtual worlds development and standardisation."[116]

Multistakeholder dialogues: Multilateral exchange on the Metaverse must also be expanded, to ensure alignment on globally accepted terms, such as fundamental human rights. A good starting point may be an expanded exchange within the US-EU Trade and Technology Council (TTC) to ensure transatlantic alignment. The World Economic Forum, partnering with INTERPOL, Meta, Microsoft and other stakeholders from academia, civil society, government and industry, launched the Defining and Building the Metaverse Initiative,[117] which includes a governance[118] and value creation track.[119] A report published in July 2023 recognises the current Global North dominance in this area, and emphasises the need for "worldwide collaboration among various stakeholders, including academics, regulators, policymakers and design teams, to nurture understanding of the Metaverse and establish protective measures." At the same time, it recognises that "unique risks may arise in different countries or regions or for different communities."[120] An inclusive and diverse consultation will be crucial to allow for the consideration of intersectional perspectives.

Conclusion

As the digital information space becomes more decentralised, generative and immersive, the severity and likelihood of risks of harm will also evolve. It is still too early to predict the exact changes, but some relevant trends can already be observed. Firstly, the decentralised social web (e.g., Odysee, PeerTube) offers a new unregulated and interconnected place of refuge and funding for malign actors. Secondly, LLMs underpinning AI applications (e.g., GPT-4.0, DALL-E) may not only generate inherent harmful content but they can potentially be exploited by malign actors for cheap, automated, and persuasive information manipulation. And thirdly, immersive worlds (e.g., Horizon Worlds) not only provide malign actors with new opportunities for mobilisation, funding, and planning, but they also potentially amplify the impact of harmful content and behaviours. This is especially the case when immersive worlds become accessible for all our senses (e.g., touch). Given the significant risks of harm, policymakers and regulators should shape the future of the analysed emerging platforms and technologies early on. The questions of liability, safety, transparency, and enforcement must be discussed anew. In the context of the convergence of new platforms and technologies, they must work closely together with academia, civil society, and industry from different sectors to gain a holistic and deep understanding of the evolving threat landscape. Moreover, they will need to continuously adapt their initiatives to mitigate the respective risks of harm and seek ways to enforce already existing policies.

Endnotes

1 **Yu, M., 2022**. *SPC Releases Typical Cases on Protection of Personality Rights*. [Online] Available at: https://www.chinajusticeobserver.com/a/spc-releases-typical-cases-on-protection-of-personality-rights [Accessed 31 July 2023]

2 **McLuhan, M., Hutchon, K., McLuhan, E., 1980**. Media, message, and language. National *Textbook Company. Skokie, IL*.

3 **World Economic Forum (WEF), 2023**. *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*. Available at: https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf [Accessed 24 August 2023]

4 **ISD, 2023**. *Digital Policy Lab*. [Online] Available at: https://www.isdglobal.org/digital-policy-lab/ [Accessed 29 August 2023]

5 **Heath, A., 2023**. *This is what Instagram's upcoming Twitter competitor looks like*. [Online] Available at: https://www.theverge.com/2023/6/8/23754304/instagram-meta-twitter-competitor-threads-activitypub [Accessed 8 August 2023]

6 **Edelman, G**. *The Father of Web3 Wants You to Trust Less*. [Online] Available at: https://www.wired.com/story/web3-gavin-wood-interview/ [Accessed 31 July 2023]

7 **Wilson, J., 2023**. *Extremist-friendly tech company closes after legal fine, The Guardian*, 16 July. Available at: https://www.theguardian.com/technology/2023/jul/16/lbry-closes-odysee-cryptocurrency-tech-sec-fraud-extremist [Accessed 8 August 2023]

8 **European Data Protection Supervisor (EDPS), 2022**. *EDPS launches pilot phase of two social media platforms*. [Online] Available at: https://edps.europa.eu/press-publications/press-news/press-releases/2022/edps-launches-pilot-phase-two-social-media_en [Accessed 31 July 2023]

9 https://social.bund.de/about

10 **TechAgainstTerrorism, 2023**. *State of play. Trends in terrorist and violent extremist use of the internet. 2022*. [Online] Available at: https://www.techagainstterrorism.org/wp-content/uploads/2023/01/FINAL-State-of-Play-2022-TAT.pdf [Accessed 8 August 2023]

11 **Christchurch Call, 2022**. *Christchurch Call Initiative on Algorithmic Outcomes*. [Online] Available at: https://www.christchurchcall.com/media-and-resources/news-and-updates/christchurch-call-initiative-on-algorithmic-outcomes/ [Accessed 8 August 2023]

12 **Hammer, D., Gerster, L. & Schwieter, C., 2023**. *Im digitalen Labyrinth. Rechtsextreme Strategien der Dezentralisierung im Netz und mögliche Gegenmaßnahmen*. [Online] Available at: https://isdgermany.org/im-digitalen-labyrinth/ [Accessed 31 July 2023]

13 **Ibid.**

14 **Gerster, A., Arcostanzo, F., Prieto-Chavana, N., et al., 2023**. *The Hydra on the Web: Challenges Associated with Extremist Use of the Fediverse – A Case Study of PeerTube*. [Online] Available at: https://www.isdglobal.org/isd-publications/the-hydra-on-the-web-challenges-associated-with-extremist-use-of-the-fediverse-a-case-study-of-peertube/ [Accessed 29 August 2023]

15 **Rozenshtein, A. Z., 2023**. *Moderating the Fediverse. Content moderation on distributed social media*. [Online] Available at: https://www.journaloffreespeechlaw.org/rozenshtein2.pdf [Accessed 31 July 2023]

16 **Kessels, B., 2022**. *The Fediverse never Forgets*. [Online] Available at: https://berk.es/2022/12/23/fediverse-never-forgets/ [Accessed 31 July 2023]

17 **Hamari, J., Koivisto, J., & Sarsa, H., 2014**. Does gamification work? A literature review of empirical studies on gamification. *2014 47th Hawaii international conference on system sciences*. 3025-3034.

18 **Matlach, P., Hammer, D. & Schwieter, C., 2022**. *On Odysee: The Role of Blockchain Technology for Monetisation in the Far-Right Online Milieu*. [Online] Available at: https://www.isdglobal.org/isd-publications/on-odysee-the-role-of-blockchain-technology-for-monetisation-in-the-far-right-online-milieu/ [Accessed 31 July 2023]

19 **Ibid.**

20 **Wilson, J., 2023**. *Extremist-friendly tech company closes after legal fine, The Guardian*, 16 July. Available at: https://www.theguardian.com/technology/2023/jul/16/lbry-closes-odysee-cryptocurrency-tech-sec-fraud-extremist [Accessed 8 August 2023]

21 **Odysee, 2022**. *Declaration of Indifference: Community Guidelines*. [Online] Available at: https://help.odysee.tv/communityguidelines/ [Accessed 29 August 2023]

22 **Hammer, D., Gerster, L. & Schwieter, C., 2023**. *Im digitalen Labyrinth. Rechtsextreme Strategien der Dezentralisierung im Netz und mögliche Gegenmaßnahmen*. [Online] Available at: https://isdgermany.org/im-digitalen-labyrinth/ [Accessed 31 July 2023]

23 **Bundesamt für Justiz (BfJ), 2023**. *Änderung der Pressemitteilung vom 2. März 2023: Bundesamt für Justiz hält Bußgeldbescheide in Höhe von 5,125 Millionen Euro gegen das soziale Netzwerk Telegram aufrecht*. [Online] Available at: https://www.bundesjustizamt.de/DE/ServiceGSB/Presse/Pressemitteilungen/2023/20230302.html [Accessed 31 July 2023]

24 **Framasoft, 2023**. *PeerTube instances*. [Online] Available at: https://instances.joinpeertube.org/instances [Accessed 29 August 2023]

25 **Future of Life Institute, 2023**. *Pause Giant AI Experiments: An Open Letter.* [Online] Available at: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ [Accessed 10 July 2023]

26 **Coulter, M., 2023**. *AI experts disown Musk-backed campaign citing their research, Reuters, 5 April.* Available at: https://www.reuters.com/technology/ai-experts-disown-musk-backed-campaign-citing-their-research-2023-03-31/ [Accessed 24 August 2023]

27 **IBM, 2023.** *What is a neural network?* [Online] Available at: https://www.ibm.com/topics/neural-networks#:~:text=Artificial%20neural%20networks%20(ANNs)%20are,an%20associated%20weight%20and%20threshold [Accessed 8 August 2023]

28 **Weidinger, L., Mellor, J., Rauh, M, et al., 2021**. *Ethical and social risks of harm from.* [Online] Available at: https://www.deepmind.com/publications/ethical-and-social-risks-of-harm-from-language-models [Accessed 10 July 2023]; Goldstein, J. A. et al., 2023. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

29 **European Parliament, 2023**. *EU AI Act: first regulation on artificial intelligence.* [Online] Available at: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence?&at_campaign=20226-Digital&at_medium=Google_Ads&at_platform=Search&at_creation=RSA&at_goal=TR_G&at_advertiser=Webcomm&at_audien [Accessed 17 July 2023]; **eSafety Commissioner, 2023**. *Tech Trends Position Statement. Generative AI.* [Online] Available at: https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai [Accessed 24 August 2023]

30 **Europol, 2023**. *ChatGPT - The impact of Large Language Models on Law Enforcement, a Tech Watch Flash Report from the Europol Innovation Lab.* Luxemburg: Publications Office of the European Union.

31 **Solaiman, I., Brundage, M., Clark, J., 2019**. *OpenAI Report: Release Strategies and the Release Strategies and the.* [Online] Available at: https://arxiv.org/ftp/arxiv/papers/1908/1908.09203.pdf [Accessed 12 July 2023]; Partnership on AI, 2023. *PAI's Responsible Practices for Practices for A Framework for Collective Action.* [Online] Available at: https://partnershiponai.org/wp-content/uploads/2023/02/PAI_synthetic_media_framework.pdf [Accessed 12 July 2023]

32 **Weidinger, L., Mellor, J., Rauh, M, et al., 2021**. *Ethical and social risks of harm from.* [Online] Available at: https://www.deepmind.com/publications/ethical-and-social-risks-of-harm-from-language-models [Accessed 10 July 2023]

33 **Partnership on AI (PAI), 2023**. *PAI's Responsible Practices for Practices for A Framework for Collective Action.* [Online] Available at: https://partnershiponai.org/wp-content/uploads/2023/02/PAI_synthetic_media_framework.pdf [Accessed 12 July 2023]

34 **Goldstein, J. A., Sastry, G., Musser, M., et al., 2023**. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

35 **Bowman, S. R., 2023**. *Eight Things to Know about Large Language Models.* [Online] Available at: https://arxiv.org/abs/2304.00612 [Accessed 31 July 2023]

36 **Gehman, S., Gururangan, S., Sap, M., et al., 2020**. *RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models.* [Online] Available at: https://arxiv.org/abs/2009.11462 [Accessed 11 7 2023]

37 **Abid, A., Farooqi , M. & Zou, J., 2021**. Large language models associate Muslims with violence. *Nat Mach Intell,* Volume 3, 461–463.

38 **Deshpande, A., Murahari, V., Rajpurohit, T., et al., 2023**. *Toxicity in CHATGPT: Analyzing Persona-assigned Language Models.* [Online] Available at: https://arxiv.org/pdf/2304.05335.pdf [Accessed 11 July 2023]

39 **Rae, J. W., Borgeaud, S., Cai, T., et al., 2021**. *Scaling Language Models: Methods, Analysis & Insights from Training Gopher.* [Online] Available at: https://arxiv.org/abs/2112.11446 [Accessed 17 July 2023]

40 **Brewster, J., Arvanitis, L. & Sadeghi, M., 2023**. *The Next Great Misinformation Superspreader: How ChatGPT Could Spread Toxic Misinformation At Unprecedented Scale.* [Online] Available at: https://www.newsguardtech.com/misinformation-monitor/jan-2023/ [Accessed 11 July 2023]

41 **Arvantis, L., Sadeghi, M. & Brewster, J., 2023**. *Despite OpenAI's Promises, the Company's New AI Tool Produces Misinformation More Frequently, and More Persuasively, than its Predecessor.* [Online] Available at: https://www.newsguardtech.com/misinformation-monitor/march-2023/ [Accessed 11 July 2023]

42 **Horvitz, E., 2023**. *On the Horizon: Interacive and Compositional Deepfakes.* [Online] Available at: https://arxiv.org/abs/2209.01714 [Accessed 31 July 2023]

43 **EEAS, 2023**. *1st EEAS Report on Foreign Information Manipulation and Interference Threats.* [Online] Available at: https://www.eeas.europa.eu/sites/default/files/documents/2023/EEAS-DataTeam-ThreatReport-2023..pdf [Accessed 11 July 2023]

44 **Goldstein, J. A., Sastry, G., Musser, M., et al., 2023**. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

45 **Van Sant, S. & Goujard, C., 2022**. *European Parliament website hit by cyberattack after Russian terrorism vote.* [Online] Available at: https://www.politico.eu/article/cyber-attack-european-parliament-website-after-russian-terrorism/ [Accessed 31 July 2023]

46 **Europol, 2023.** *ChatGPT - The impact of Large Language Models on Law Enforcement, a Tech Watch Flash Report from the Europol Innovation Lab.* Luxemburg: Publications Office of the European Union.

47 **Hill, M., 2023**. *6 ways generative AI chatbots and LLMs can enhance cybersecurity (CSO).* [Online] Available at: https://www.csoonline.com/article/575377/6-ways-generative-ai-chatbots-and-llms-can-enhance-cybersecurity.html [Accessed 31 July 2023]

48 **Bessi, A. & Emilio, F., 2016**. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday,* 21(11); Wang, E. L., Luceri, L., Pierri, . F. & Ferrara, E., 2023. *Identifying and characterizing behavioral classes of radicalization within the QAnon conspiracy on Twitter.* s.l., Proceedings of the International AAAI Conference.

49 **Heubl, B., Sabolwski, N. & Weinmann, L., 2023.** Vulkan Files #michgibtsgarnicht, SZ, 31 March. Available at: https://www.sueddeutsche.de/projekte/artikel/politik/russland-cyberkrieg-desinformation-propaganda-fakenews-twitter-vulkan-files-ukraine-e287057/?reduced=true [Accessed 31 July 2023]

50 **Europol, 2023.** *ChatGPT - The impact of Large Language Models on Law Enforcement, a Tech Watch Flash Report from the Europol Innovation Lab.* Luxemburg: Publications Office of the European Union.

51 **Goldstein, J. A., Sastry, G., Musser, M., et al., 2023**. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

52 **Ibid.**

53 **Meta Fundamental AI Research Diplomacy Team (FAIR), Brown, N., Dinan, E., et al., 2022**. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science,* 378(6624).

54 **Bai, H., Voelkel, J. G., Eichstaedt, J. C., et al., 2023**. *Artificial Intelligence Can Persuade Humans on Political Issues.* [Online] Available at: https://osf.io/stakv/ [Accessed 17 July 2023]

55 **Spitale, G., Biller-Andorno, N. & Germani, F., 2023**. AI model GPT-3 (dis)informs us better than humans. *ScienceAdvances,* 9(26).

56 **Weidinger, L., Mellor, J., Rauh, M, et al., 2021**. *Ethical and social risks of harm from.* [Online] Available at: https://www.deepmind.com/publications/ethical-and-social-risks-of-harm-from-language-models [Accessed 10 July 2023]

57 **YouGov, 2023**. *KI – Chance oder Bedrohung?.* [Online] Available at: https://yougov.de/topics/technology/articles-reports/2023/05/17/ki-chance-oder-bedrohung [Accessed 17 July 2023]

58 **Etienne, H., 2021**. The future of online trust (and why Deepfake is advancing it). *AI Ethics,* 1, 553-562.

59 **Anderljung, M., Barnhart, J., Leung, J., et al., 2023**. *Frontier AI Regulation: Managing Emerging Risks to Public Safety.* [Online] Available at: arxiv.org/pdf/2307.03718.pdf [Accessed 26 July 2023]

60 **CSIS, 2023**. *Sen. Chuck Schumer Launches SAFE Innovation in the AI Age at CSIS.* [Online] Available at: https://www.csis.org/analysis/sen-chuck-schumer-launches-safe-innovation-ai-age-csis [Accessed 31 July 2023]

61 **White House, 2023**. *Ensuring Safe, Secure, and Trustworthy AI.* [Online] Available at: Ensuring-Safe-Secure-and-Trustworthy-AI.pdf (whitehouse.gov) [Accessed 26 July 2023]

62 **European Union, 2023**. *MEPs ready to negotiate first-ever rules for safe and transparent AI.* [Online] Available at: https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai [Accessed 26 July 2023]

63 **European Commission, 2023**. *Artificial intelligence: in Europe, innovation and safety go hand in hand. Statement by Commissioner Thierry Breton.* [Online] Available at: Artificial intelligence | Statement by Commissioner Breton (europa.eu) [Accessed 26 July 2023]

64 **Novelli, C., Taddeo, M. & Floridi, L., 2023**. *Accountability in artificial intelligence: what it is and how it works.* [Online] Available at: https://link.springer.com/article/10.1007/s00146-023-01635-y [Accessed 31 July 2023]

65 **Ibid.**

66 **Ariyaratne, H., 2023**. *ChatGPT and intermediary liability: Why Section 230 does not and should not protect generative algorithms.* [Online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4422583 [Accessed 31 July 2023]

67 **eSafety Commissioner, 2023**. *Tech Trends Position Statement. Generative AI.* [Online] Available at: https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai [Accessed 24 August 2023]

68 **Goldstein, J. A., Sastry, G., Musser, M., et al., 2023**. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

69 **Buchanan, B., Lohn, A., Musser, M., et al., 2021**. *Truth, Lies, and Automation. How language models could change disinformation.* Center for Security and Emerging Technology, Washington, DC.

70 **Deshpande, A., Murahari, V., Rajpurohit, T., et al., 2023**. *Toxicity in CHATGPT: Analyzing Persona-assigned Language Models.* [Online] Available at: https://arxiv.org/pdf/2304.05335.pdf [Accessed 11 July 2023]

71 **Greer, C., 2017**. *International Personality Rights and Holographic Portrayals.* [Online] Available at: https://doi.org/10.18060/7909.0052 [Accessed 26 July 2023]

72 **Celli, F., 2020**. *Deepfakes Are Coming: Does Australia Come Prepared?* [Online] Available at: http://classic.austlii.edu.au/au/journals/CanLawRw/2020/18.pdf [Accessed 26 July 2023]

73 **eSafety Commissioner, 2023**. *Tech Trends Position Statement. Generative AI.* [Online] Available at: https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai [Accessed 24 August 2023]

74 **Goujard, C., 2023**. EU wants Google, Facebook to start labeling AI-generated content. *POLITICO*. Available at: https://www.politico.eu/article/chatgpt-dalle-google-facebook-microsoft-eu-wants-to-start-labeling-ai-generated-content/ [Accessed 29 August 2023]

75 **Goldstein, J. A., Sastry, G., Musser, M., et al., 2023**. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

76 **Coalition for Content Provenance and Authenticity (C2PA)**. *Overview.* [Online] Available at: https://c2pa.org/ [Accessed 8 August 2023]

77 **Partnership on AI (PAI)**, 2023. *About Us.* [Online] Available at: https://partnershiponai.org/about/ [Accessed 8 August 2023]

78 **Anderson, J. & Rainie, L., 2023**. *The Metaverse in 2040.* [Online] https://www.pewresearch.org/internet/2022/06/30/the-metaverse-in-2040/ [Accessed 31 July 2023]

79 **Ibid.**

80 **The New York Times, 2021.** The Metaverse Is Mark Zuckerberg's Escape Hatch, *The New York* Times, 29 October. Available at: https://www.nytimes.com/2021/10/29/technology/meta-facebookzuckerberg.html [Accessed 31 July 2023]

81 **Meta, 2021**. *Connect 2021: Our vision for the metaverse.* [Online] Available at: https://tech.fb.com/ar-vr/2021/10/connect-2021-our-vision-for-the-metaverse/ [Accessed 31 July 2023]

82 **Newton, C., 2021**. *Mark in the metaverse.* [Online] Available at: https://www.theverge.com/22588022/mark-zuckerberg-facebook-ceo-metaverse-interview [Accessed 31 July 2023]

83 **Ball, M., 2020**. *The Metaverse: What It Is, Where to Find it, and Who Will Build It.* [Online] Available at: https://www.matthewball.vc/all/themetaverse [Accessed 31 July 2023]

84 **Heath, A., 2021**. Meta opens up access to its VR social platform Horizon Worlds, *The Verge*, 9 December. Available at: https://www.theverge.com/2021/12/9/22825139/meta-horizon-worlds-access-open-metaverse [Accessed 31 July 2023]

85 **World Economic Forum (WEF), 2023**. Interoperability in the Metaverse. [Online] Available at: https://www.weforum.org/reports/interoperability-in-the-metaverse/ [Accessed 24 August 2023]

86 **Europol, 2022**. *Policing in the metaverse: what law enforcement needs to know.* Luxembourg: Publications Office of the European Union.

87 **Council of the European Union, 2022**. *The Metaverse in the context of the fight against terrorism.* EU Counter-Terrorism Coordinator. [Online] Available at: https://data.consilium.europa.eu/doc/document/ST-9292-2022-INIT/en/pdf [Accessed 8 August 2023]

88 **WeProtect Global Alliance, 2023**. Intelligence briefing. Extended Reality technologies and child sexual exploitation and abuse. [Online] Available at: https://www.weprotect.org/library/extended-reality-technologies-and-child-sexual-exploitation-and-abuse/ [Accessed 24 August 2023]

89 **World Economic Forum (WEF), 2023**. Metaverse Privacy and Safety. [Online] Available at: https://www3.weforum.org/docs/WEF_Metaverse_Privacy_and_Safety_2023.pdf [Accessed 24 August 2023]

90 **Ashraf, M. 2023**. *Gender-based Abuse on the Metaverse: The New Internet is Being Coded on a Toxic Palimpsest.* [Online] Available at: https://botpopuli.net/gender-based-abuse-on-the-metaverse-the-new-internet-is-being-coded-on-a-toxic-palimpsest/ [Accessed 31 July 2023]

91 **Patel, N. J., 2021**. *Reality or Fiction?* [Online] Available at: https://medium.com/kabuni/fiction-vs-non-fiction-98aa0098f3b0 [Accessed 31 July 2023]

92 **Gray, K.L, 2016**. Solidarity is for white women in gaming. Using Critical Discourse Analysis To Examine Gendered Alliances and Racialized Discords in an Online Gaming Forum. In: Kafai, Y.B., Richard, G.T., Tynes, B.M., et al., 2016. *Diversifying Barbie and Mortal Kombat: Intersectional perspectives and inclusive designs in gaming.* Carnegie Mellon University, ETC Press, Pittsburgh, PA.

93 **WeProtect Global Alliance, 2023**. Intelligence briefing. Extended Reality technologies and child sexual exploitation and abuse. [Online] Available at: https://www.weprotect.org/library/extended-reality-technologies-and-child-sexual-exploitation-and-abuse/ [Accessed 24 August 2023]

94 **Blackwell, L., Ellison, N., Elliott-Deflo, N., & et al., 2019**. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction Vol. 3* Issue CSCW, Article No.: 100,1–25.

95 **Mrug, S., Madan, A. & Windle, M., 2016**. Emotional Desensitization to Violence Contributes to Adolescents' Violent Behavior. *Journal of Abnormal Child Psychology*, 44 (1), 75-86.

96 **Rankin C.H. et al., 2009**. Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. *Neurobiology of Learning and Memory*. 92, 135-138.

97 **Mrug, S., Madan, A. & Windle, M., 2016**. Emotional Desensitization to Violence Contributes to Adolescents' Violent Behavior. *Journal of Abnormal Child Psychology*, 44 (1), 75-86.

98 **Hayward, K.J. & Cottee, S., 2011**. Terrorist (e)motives: the existential attractions of terrorism. *Studies in Conflict and Terrorism*, 34(12), 963-986.

99 **Council of the European Union, 2022**. *The Metaverse in the context of the fight against terrorism*. EU Counter-Terrorism Coordinator. [Online] Available at: https://data.consilium.europa.eu/doc/document/ST-9292-2022-INIT/en/pdf [Accessed 8 August 2023]

100 **Europol, 2022**. *Policing in the metaverse: what law enforcement needs to know*. Publications Office of the European Union, Luxemburg.

101 **The Algemeiner, 2022**. *Children's Gaming Platform Removes 'Disturbing' Nazi Concentration Camp 'Experience' With Gas Chambers*. [Online] Available at: https://www.algemeiner.com/2022/02/21/childrens-gaming-platform-removes-disturbing-nazi-concentration-camp-experience-with-gas-chambers/ [Accessed 31 July 2023]

102 **Financial Times**. *NFTs: The metaverse economy*. [Online] Available at: https://www.ft.com/partnercontent/crypto-com/nfts-the-metaverse-economy.html [Accessed 31 July 2023]

103 **Council of the European Union, 2022**. *The Metaverse in the context of the fight against terrorism*. EU Counter-Terrorism Coordinator. [Online] Available at: https://data.consilium.europa.eu/doc/document/ST-9292-2022-INIT/en/pdf [Accessed 8 August 2023]

104 **World Economic Forum, 2023**. *Metaverse Privacy and Safety*. [Online] Available at: https://www3.weforum.org/docs/WEF_Metaverse_Privacy_and_Safety_2023.pdf [Accessed 31 July 2023]

105 **Europol, 2022**. *Policing in the metaverse: what law enforcement needs to know*. Publications Office of the European Union, Luxemburg.

106 **World Economic Forum, 2023**. *Metaverse Privacy and Safety*. [Online] Available at: https://www3.weforum.org/docs/WEF_Metaverse_Privacy_and_Safety_2023.pdf [Accessed 31 July 2023]

107 **Ibid.**

108 **Tuck, H., Guhl, J., et al., 2023**. *Researching the Evolving Online Ecosystem: Telegram, Discord & Odysee*. [Online] Available at: https://www.isdglobal.org/isd-publications/researching-evolving-online-ecosystem-telegram-discord-odysee/ [Accessed 8 August 2023]

109 **World Economic Forum, 2023**. *Metaverse Privacy and Safety*. [Online] Available at: https://www3.weforum.org/docs/WEF_Metaverse_Privacy_and_Safety_2023.pdf [Accessed 31 July 2023]

110 **Meta, 2023**. *What is the Safe Zone in Horizon Worlds?* [Online] Available at: https://www.meta.com/en-gb/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/safe-zone-in-horizon/ [Accessed 31 July 2023]

111 **Meta, 2022**. *Introducing a Personal Boundary for Horizon Worlds and Venues*. [Online] Available at: https://about.fb.com/news/2022/02/personal-boundary-horizon/ [Accessed 31 July 2023]

112 **Blackwell, L., Ellison, N. Elliott-Deflo, N., & Schwartz, R., 2019**. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), Art. No. 100, 1-25.

113 **Chandrasekharan, E. et al., 2018**. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer* Interaction, 2(CSCW), Art. No. 32, 1–25.

114 **Blackwell, L., Ellison, N. Elliott-Deflo, N., & Schwartz, R., 2019**. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), Art. No. 100, 1-25.

115 **European Commission, 2023**. *Virtual worlds (metaverses) – a vision for openness, safety and respect*. [Online] Available at: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13757-Virtual-worlds-metaverses-a-vision-for-openness-safety-and-respect_en [Accessed 31 July 2023]

116 **Ibid.**

117 **World Economic Forum, 2023**. *Partners*. [Online] Available at: https://initiatives.weforum.org/defining-and-building-the-metaverse/partners [Accessed 31 July 2023]

118 **Ibid.**

119 **Ibid.**

120 **World Economic Forum, 2023**. *Metaverse Privacy and Safety*. [Online] Available at: https://www3.weforum.org/docs/WEF_Metaverse_Privacy_and_Safety_2023.pdf [Accessed 31 July 2023]

ISD | Institute for Strategic Dialogue

Amman | Berlin | London | Paris | Washington DC

www.isdgermany.org

Sponsored by:

Federal Foreign Office