REPHRAIN

Protecting citizens online

REPHRAIN White Paper: the Metaverse and Web 3.0

The following researchers contributed to this report (in alphabetical order): Dr Aydin Abadi, Prof Madeline Carr, Dr Ignacio Castro, Dr Alicia Cork, Dr Andrés Domínguez, Cristina Fiani, Dr Mohamed Khamis, Dr Mark McGill, Prof Steven Murdoch, Prof Awais Rashid, Dr Pejman Saeghe, Dr Gareth Tyson.

The submission was edited by Dr Ola Michalec (Policy Engagement Associate).

May 2023

REPHRAIN White Paper: the Metaverse and Web 3.0

REPHRAIN National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online

May 2023

Introduction

Social, political and ethical implications of Metaverse and Web 3.0 are a growing concern as the use of virtual worlds and decentralised web solutions continues to increase. In particular, the potential for online harms requires further academic inquiry as well as adequate regulatory measures. Online harms can manifest in a variety of forms, ranging from cyberbullying and trolling to violations such as the sexual exploitation of children, intellectual property theft, and unauthorised access to personal data.

This report synthesises the outputs from the REPHRAIN Research Centre on Privacy, Harm Reduction and Adversarial Influence Online to highlight online harms in the Metaverse and Web 3.0 and discuss ways to mitigate and regulate them. The report presents both research findings as well as recommendations to policymakers and developers of emerging technologies. The target audience of the report are regulators of digital technologies, developers of immersive and decentralised solutions, civic rights groups and interested researchers.

Our report is divided into the following sections[1]. Sections 1-4 present research landscape, while section 5 offers recommendations. In section 1, we outline evidence on cyber security and data protection, highlighting vulnerabilities and potential solutions. In section 2, we specify harms pertaining to the use of immersive and Web 3.0 technologies. In section 3, we argue for a responsible innovation research policy, which is cognizant of improving users' agency, literacy, and accessibility. In section 4, we bring our concerns regarding appropriate regulatory frameworks and standardisation initiatives. Finally, section 5 lists policy recommendations for robust and responsible regulation of both immersive and Web 3.0 technologies.

About REPHRAIN

[1] This report is based on the REPHRAIN's recent contribution to the All-Party Parliamentary Group on the Metaverse and Web 3.0, which can be accessed here: https://bpb-eu-w2.wpmucdn.com/blogs.bristol.ac.uk/dist/1/670/files/2023/03/Call-for-Papers-by-the-All-Party-Parliamentary-Group-REPHRAIN-Response.pdf

REPHRAIN, the National **Re**search Centre on **P**rivacy, **H**arm **R**eduction and **A**dversarial **In**fluence O**n**line, is the UK's world-leading interdisciplinary community focused on the protection of citizens online. As a UKRI-funded National Research Centre, we boast a critical mass of over 100 internationally leading experts at 13 UK institutions working across 37 diverse research projects and 23 founding industry, non-profit, government, law, regulation and international research centre partners. As an interdisciplinary and engaged research group, we work collaboratively on addressing the three following missions:

- Delivering privacy at scale while mitigating its misuse to inflict harms
- Minimising harms while maximising benefits from a sharing-driven digital economy
- Balancing individual agency vs. social good.

The REPHRAIN Centre has extensive expertise in identifying, contextualising, and mitigating online harms in the Web 3.0 and Metaverse environments. We are currently working on the following projects concerned with the Metaverse and Web 3.0 technologies:

- Project PriXR intends to futureproof Extended Reality (XR[2]) technology against violations of privacy and anonymity. It explores XR not only in terms of its benefits to society, but also in supporting resistance against surveillance and misuse and facilitating bystander awareness and consent, so that we can safely unlock the benefits of this emerging technology.
- The EMBODY project explores the potential psychological harms that may arise through the widespread adoption of virtual reality (VR) technologies.
- The HARM project provides a framework for categorising and anticipating online and virtual reality harms.
- Project VIRRAC aims to understand the extent of child abuse in the Metaverse environments to raise awareness, develop policy guidance and tools for law enforcement
- The MetaSafeChild project aims to assess the potential harms faced by children and adolescents in social Virtual Reality (VR) platforms and develop user-centered moderation tools for in-VR conflicts.
- Project PAYMENT is developing blockchain-based protocols as privacy preserving mechanisms to support the fair exchange of digital coins.
- Project Cryptocurrency Fraud is developing a decentralised fraud recovery mechanism in the context of digital currencies
- Project DSNmod aims to tackle challenges of decentralised social networks by minimising online harms and exploring privacy-preserving federated moderation.
- In addition, we have a dedicated group of researchers who focus on Responsible Innovation. To anticipate ethical and social issues, they conduct discourse analysis of the future claims and responsible innovation pledges by technology companies, as well as a range of publications by industry analysts and the media related to past controversies, emerging technical developments, and future trends around the Metaverse.
- Finally, REPHRAIN Policy and Regulation strand synthesises research outputs and translates them into actionable policy evidence in the form of reports, events, consultations and research collaborations.

[2] Extended Reality (XR) is an umbrella term for immersive technologies like Virtual Reality (VR), Augmented Reality (AR) and Mixed Reality (MR).

The following researchers contributed to this report (in alphabetical order): Dr Aydin Abadi, Prof Madeline Carr, Dr Ignacio Castro, Dr Alicia Cork, Dr Andrés Domínguez, Cristina Fiani, Dr Mohamed Khamis, Dr Mark McGill, Prof Steven Murdoch, Prof Awais Rashid, Dr Pejman Saeghe, Dr Gareth Tyson. The submission was edited by Dr Ola Michalec (Policy Engagement Associate).

Please cite this resource as: Michalec, O., Abadi A., Carr, M., Castro, I., Cork, A., Domínguez, A., Fiani, C., Khamis, M., McGill, M., Murdoch, S., Rashid, A., Saeghe, P., Tyson, G., (2023) REPHRAIN White Paper: the Metaverse and Web 3.0. Report.

1. Cybersecurity & data protection

Usable and Secure Authentication in XR

There is a growing need for secure and usable authentication for Extended Reality (XR). So far, existing platforms have been using established authentication schemes, like PINs, passwords, and Lock Patterns to secure access to devices, verify identity, and confirm purchases. However, existing methods suffer from security and usability problems. For example, users are required to come up with many, ideally unique, passwords that feature digits and symbols, but humans are cognitively unable to remember these many passwords. Authentication takes time and is perceived by users as an obstacle, which prompts them to work around it by writing their passwords down or using weak easy-to-remember passwords. On the security side, XR headsets introduce a unique threat in that user's input during authentication can be observed by bystanders as shown in the figure on the right.

Our recent research from PriXR project highlights the potential for utilising XR technology in developing more usable and secure authentication methods. One potential approach is the use of behavioural biometrics, where a user's unique behaviour is used to identify them. This is particularly relevant in XR environments, as the 3D spaces provide a wide range of possibilities for user movements [1]. There is also a potential for using 3D objects and environments as a means of authentication. This can significantly expand the password space and facilitate faster user authentication [2]. While biometrics-based authentication methods show potential for improved security, they pose significant privacy implications (these are discussed in the Section 2).

Dealing with financial cybercrime with the help of Web 3.0 technologies

Web 3.0 technologies show a potential to tackle financial cybercrime. An example of this is the research from PAYMENT project exploring "Authorised Push Payment" fraud resolution. An "Authorised Push Payment" (APP) fraud is a type of cybercrime where a fraudster tricks a victim into making an authorised online payment into an account controlled by the fraudster. According to a report produced by "UK Finance" [3], only in the first half of 2021, a total of £355 million was lost to APP frauds, which has increased by 71% compared to losses reported in the same period in 2020.

Although the amount of money lost via APP fraud and the number of cases has been significantly increasing, the victims are not receiving enough protection. In the first half of 2021, only 42% of the stolen funds were returned to victims of APP fraud in the UK. Despite the UK's financial regulators having provided a specific reimbursement framework (I.e., Contingent Reimbursement Model (CRM) code [4]), this framework is currently voluntary and open to interpretation. Furthermore, there exists no transparent and uniform mechanism via which honest victims can prove their innocence.

To facilitate the compensation of APP frauds victims for their losses, we proposed a blockchain-based protocol[3] called "Payment with Dispute Resolution" and proved the protocol's security [5, 6]. The protocol lets an honest fraud victim independently prove its innocence to a third-party dispute resolver, to be reimbursed. The protocol makes use of a standard online banking system; hence it does not require significant changes to the existing online banking systems. It also automates the implementation of reimbursement regulations where possible.

The protocol demonstrates promising capabilities of Web 3.0 technologies such as (i) accurately formalising reimbursements' conditions (ii) offering traceability: it lets parties' performance be tracked, and (iii) providing an evidence-based final decision: it requires the reasons leading to the final decision to be accessible and consistent with the reimbursements' conditions and parties' actions. It also offers accountability, as it is equipped with auditing mechanisms that help identify the party liable for an APP fraud loss. We hope that our result lays the foundation for future solutions that will protect victims of this concerning type of fraud.

Security and Privacy of cryptocurrency payment mechanisms

As the Internet's use for conducting business is rapidly growing, it requires the provision of secure payment mechanisms. This problem can be illustrated with a variety of real-world scenarios; for instance, when two parties want to exchange digital items or when a seller wants to sell a digital verifiable service in exchange for cryptocurrencies. Solutions to the problem are usually certain cryptographic schemes, called fair exchange protocols, and have been studied for decades.

With the advent of decentralised cryptocurrencies and blockchain, fair exchange protocols which do not rely on a single trusted third party (e.g., a bank) have been proposed. In such a scenario, the third party's role can be turned into a computer program, a so-called smart contract, which is maintained and executed by the decentralised blockchain. In the future, this could ultimately result in a stronger security guarantee, as there would be no need to trust a single entity anymore.

However, **our findings from the PAYMENT project [5, 6] highlight security and privacy issues with previously proposed fair exchange protocols** [7]. We identified a free-riding attack that lets an adversary use a service without paying the fee. We also found that the protocol leak information about the seller and a buyer to the public, e.g., deposits' amounts and proof status.

[3] Blockchain protocol refers to a set of rules define the interface of the network, interaction between the computers, incentives, kind of data, etc. They establish the structure of the blockchain — the distributed database that allows digital money to be securely exchanged on the internet.

We, therefore, proposed the improvement called “Recurring Contingent Service Payment” which mitigates the attack and preserves the parties’ privacy.

2. Safeguarding users & non-users, particularly children & most vulnerable

The typology of harms pertaining to immersive technologies

The REPHRAIN Centre researchers have been developing typologies of online harms pertaining to immersive technologies and beyond. Here, the Centre’s flagship resource, “The Map of Online Harms, Risks, and Vulnerabilities [8], defines and describes 15 harms, linking them to 6 positive attributes (financial security, reputation, fairness, safety and wellbeing, privacy, freedom of speech). The Map signposts to ongoing debates, research challenges and seminal resources, enabling policymakers evidence-based anticipatory governance of harms in emerging technologies.

Furthermore, recent research from the HARM project (pre-print available at: Dr Alicia Cork ac974@bath.ac.uk [9]) determined that there are six key vulnerabilities to consider when conceptualising the full spectrum of harm. We believe that this categorisation scheme may help to inform policy recommendations.

The first type of harm considers **user access to information**; a harm that arises from experiencing or viewing risky content, such as violent imagery, abuse, or pornography, a concern especially important to immersive environments which currently lack adequate safeguarding tools.

The second vulnerability relates to **user supplied information**; a harm that arises through the active behaviour of the user and the voluntary sharing or creation of content. In the context of immersive technologies, there is a regulatory gap around ensuring that virtual identities link to real-world identities.

The third vulnerability relates to **access, storage, and the collection of data**; a harm that arises from the tracking of individuals, but also unauthorised access to that data (e.g., from cyber attacks or theft). Currently XR devices lack adequate user privacy and security protections (despite holding granular personal data on, for example, biometrics), while bystanders (non-users) have no capacity to consent how/if they’re being recorded.

The fourth vulnerability relates to the **processing of data**; a harm that arises from the sale of tracking data, and the development of unethical or unregulated algorithms. The potential for XR technologies to collect realms of highly sensitive personal data cannot be overstated.

The fifth vulnerability relates to **accessibility and misuse risks**; harm relating to the exclusion of certain groups (in terms of financial cost or disability, e.g., cost of headsets or accessibility for those with visual impairment), as well as physical harms such as cybersickness (a technology-induced version of motion sickness caused by moving content on screens).

The sixth and final harm relates to **decisional interference risks;** the impact of features and content designed to manipulate and nudge behaviour (e.g., dark patterns). Examples of dark patterns include designing interfaces so users do not consider their privacy, or overloading users with information such that they cannot make informed decisions [38] (see section on XR-enabled Deceptive Design below).

This framework helps to conceptualise the different types of regulatory interventions necessary to keep citizens safe in virtual reality environments. By breaking harms down into these six types, we can begin to understand how the interplay of technological features, users and bystanders, and data collected and presented can be regulated in tandem to create a safe environment in which to enjoy the potential of metaverse technologies.

Social Virtual Reality Use Amongst Children and Adolescents

The need for robust social VR safeguarding measures to protect users' privacy and well-being (particularly for children) becomes increasingly important. Social VR is growing technology designed to facilitate social gatherings with the help of apps and headsets. The past couple of years have seen social VR reach the consumer market and grow significantly, especially with the pandemic accelerating adoption [10,11,12]. One of the mainstream social VR platforms, VRChat has an estimated 7.2 million players in total [13]. Social VR, initially designed for adults, has attracted teenagers and younger children [14, 15]. They have been drawn to social VR because of its engaging and immersive activities, allowing them to connect with friends beyond just playing games [15]. On the downside, there has been an increase in harassment, bullying and new forms of harm in social VR [16, 17]. Children and adults have been reporting experiences of harassment, from name calling to physical stalking and sexual harassment [17].

Cristina Fiani’s PhD project (2022-2025), supervised by REPHRAIN researchers, Dr Mohamed Khamis and Dr Mark McGill, focuses on safeguarding children from social VR disruptions, developing safety-enhancing tools to protect children from harassment and bullying in social VR and allowing effective parental oversight.

Our first study consisted of **evaluating parents’ and non-parents’ perspectives on children’s use of social VR** via a mixed-methods questionnaire (149 responses, including 79 parents). We collected stories of notable social VR encounters between children and adults to provide the most complete picture yet of how the presence of children in shared child/adult spaces introduces safeguarding challenges for both parties and also of how parents would choose to moderate or otherwise limit their child's usage of social VR across their childhood. Our results reveal new insights into the extent to which minors use social VR; how in our sample parenthood, familiarity, and supervision influence perceptions of children's use; and how parents would choose to moderate or to otherwise limit their child's usage of social VR across their childhood. (Pre-Print of Paper [18] available on demand, email Cristina Fiani: c.fiani.1@research.gla.ac.uk).

Our second study aimed to **measure children’s and parents’ perceptions of a Wizard-Of-Oz[4] AI-Embodied Moderator (Big Buddy)** safeguarding children from potential harassment in a simulated social VR environment. The results showed that children felt reassured and safer with the presence and intervention of Big Buddy and perceived its role as a referee. Additionally, children preferred more realistic and humanised authority figures, and familiar and positive-related figures. We hope that our findings contribute to a better understanding of children's

[4] Wizard of Oz prototyping is creating a prototype that seems to work automatically from the user’s perspective but, in reality, it’s a person controlling the prototype’s responses on the other end

perceptions towards embodied AI-moderators in social VR, and lead to future research on safer, inclusive, and personalised AI-moderators for different groups based on children's age, social VR environments, and parental oversight needs. (Pre-Print of Paper [19] also available on demand, email Cristina Fiani: c.fiani.1@research.gla.ac.uk).

Safeguarding XR Users

Extended Reality (XR) devices introduce an unprecedented ability to surveil our everyday lives, with a potential to violate our privacy. First, in our homes and offices (for instance, think of VR home entertainment and VR in the context of future work scenarios), and later in public life (particularly via AR). The large amount of data collected by XR devices is necessary for their functionality and to deliver powerful benefits to XR users (e.g., augmented intelligence and augmented perception). However, this amount of data (including personal data) collected from the users opens avenues for exploitation and abuse, putting people at risk [20].

Project PriXR has shown that beyond privacy, XR users' freedom of thought [21] and safety [22] are also at risk. VR headsets, in particular, has been shown to be capable of manipulating VR users' sense of movement, potentially putting VR users and their bystanders in physical harm's way [23]. Users are also acutely vulnerable to bystanders, due to users' immersion in VR, which reduces users' awareness of their surroundings [24]. Regarding freedom of thought, XR will enable real-time manipulation of users' thoughts, actions, and behaviours, driven by contextual understanding married with devices' ability to augment or alter our perception of reality [21]. XR devices have also been shown capable of manipulating memory [25], and we suggest the coming years will see new exploits of XR, amplifying information disorder, including mis-, dis-, and mal-information, resulting in attitudinal change [26].

Furthermore, in the EMBODY project, we are exploring how individuals may be vulnerable to misinformation – or even 'mis-experience' – in virtual reality. Whilst there is a great body of work focusing on how to protect individuals from misinformation or 'fake news' in online spheres [27], there is currently very little research which looks at the power of virtual reality to misinform individuals. **We suggest that the immersive and emotional components of virtual reality have the potential to render individuals exceptionally vulnerable to manipulation due to the sense of 'realness' that is conveyed when immersed in a virtual environment.** It is therefore imperative to better understand how we can protect individuals from this manipulation, be it through psychologically informed changes to technical design features, or through recommendations for concise legal regulations. Additionally, we are also researching the role of photorealistic avatars within this process, identifying whether individuals perceive photorealistic avatars as more trustworthy than their cartoon counterparts (pre-prints available on demand, please contact Dr Alicia Cork ac974@bath.ac.uk).

Safeguarding Non-Users / Bystanders to XR

Project PriXR has also explored safeguards for bystanders to XR experiences**. For bystanders to XR devices, mass adoption of wearables (e.g., Augmented Reality glasses) poses challenges around mass surveillance; eroding the concept of public privacy**. But beyond recording and volumetric capture, everyday AR can also reveal deeper insights into bystanders, being able to gain insights into their bystanders' mental/cognitive processes, infer their stress/arousal levels and affective states, and infer sensitive personal information and characteristics. This is made possible due to the wealth of data captured by such devices from on-board cameras and microphones [28]. Moreover, such devices might be able to unilaterally alter the bystanders' appearance and expression of social identity without their knowledge or consent, opening society up to new forms of abuse and harassment. Lemley et al. considered the legality of this ability to augment our personal sensescape and the sensescapes of others, asking: ``*What if people use this... to make [you] appear ridiculous... without your knowledge or consent? Or what if they want to make you appear naked"* [29], potentially necessitating "consent to augment" [30] or the creation of awareness and consent mechanisms that give bystanders agency over how they are sensed and augmented [29].

XR-enabled Deceptive Design

Deceptive designs (also known as dark patterns) have traditionally been used to manipulate users into behaviours that benefit the service owners, typically at the expense of the users. Think of 'disguised ads' where ads pretend to be other types of content to trick users into clicking on them, or 'confirm shaming' where the decline option is worded in a way that shames users into compliance, for instance to subscribe to a premium service. As XR devices become popular, dark patterns are likely to be used to manipulate users. **Given the characteristics of XR devices, for instance the myriad of embedded sensors and the ability to immerse the users, the impact of dark patterns will be exacerbated, and novel XR-specific dark patterns are likely to emerge.** Not mitigating dark patterns in this context will have a negative impact on users' security, safety, and privacy, in turn eroding users' trust in the underlying technologies. In the project PriXR, we are currently investigating various aspects of this phenomenon, with a paper in review examining the risks dark patterns pose in amplifying and extending the deceptive designs available to malicious attackers [31] (a copy available on demand from Mark.McGill@glasgow.ac.uk).

Content moderation in decentralised social media

Decentralisation is core to the Web 3.0. As part of the efforts to decentralise the Web*,* Decentralised Social Networks (e.g., Mastodon, Pleroma) have become popular. These services offer microblogging services similar to centralised applications like Twitter. The recent acquisition of Twitter has brought a large number of users to these platforms.

Studies from DSNmod researchers explored the challenges of content moderation in the Decentralised Web. A critical difference with Twitter is that decentralised social networks are composed of independent servers (also called instances) which are moderated by independent administrators. Users, however, can interact with each other regardless of the instance where their account is providing a similar service and perception to their centralised alternative (i.e., Twitter). **This decentralisation introduces new challenges, among which DSNmod researchers identify:** 1) a heavy moderation load on administrators [32, 33], and 2) fundamental limitations in the current moderation tools such as blocking all users from an instance rather than just the offenders [33, 34]. As a result, the researchers find widespread presence of toxic content [33, 34, 35].

The work also identifies avenues to address these challenges: 1) new streamlined user-driven policies that enable administrators to moderate on a per-user (rather than per-instance) basis in an easier and potentially semiautomated fashion [34], 2) streamlining moderation with the assistance of automated classifiers (a technique relying on machine learning to classify information as complying with moderation policy) [35]. The latter however is complicated by the decentralisation. This is because of the economies of scale of machine learning, where classifiers become more effective when trained in larger pools of data, but differently from Twitter where all content can be aggregated, in decentralised social networks, this information is scattered across instances.

In order to overcome the problem of economies of scale in decentralisation, the researchers show how **federated learning** (I.e., machine learning techniques which train an algorithm across multiple decentralised servers holding local data samples, without exchanging them) allows multiple entities to train together an algorithm without sharing the actual data. This privacy preserving approach can help reduce the barrier of entry for services where large economies of scale can deter new entrants and reduce competition. This is particularly the case for social networks, where economies of scale and network effects have frequently resulted in limited competition and monopolistic behaviours [35].

3. Responsible & sustainable innovation

Digital literacy

As the Metaverse and Web 3.0 technologies are gaining traction, we posit that significant investment will be needed in digital and futures literacy for different stakeholders including users, non-users, academics, journalists, and policy makers. These interventions ought to address security, privacy protection, safeguarding and wider ethical concerns raised by the potential mass adoption of the Web 3.0, the Metaverse and its enabling immersive technologies.

Literacy interventions should focus on improving the ability to use of new interfaces (e.g., headsets, handsets, or Web 3.0 user graphic interfaces), improving users' agency and control over their preferences, understanding of the potential risks, addressing deceptive designs and adoption experiences, and anticipating individual and social harms. These actions could be taken up by experts in digital literacy and education, human computer interaction, usability, and developer communities in collaboration with policy stakeholders like Ofcom.

Responsible innovation in the Metaverse

The recent hype around the potential mainstreaming of the Metaverse and emerging immersive technologies is raising many concerns over privacy, online safety, unfair data extraction and the reproduction of harms that have already been evidenced extensively in the realm of social media platforms. Meta's vision of the Metaverse builds on a large existing user base and a business model that trades on an economy of data and attention where users are incentivised to spend more time online and share more data. While the metaverse introduces new interfaces and immersive modes of accessing the internet, the economics of social media are still measured by user growth and engagement which drives companies to strive for market monopoly and network effects to grow their user base [36]. **In this context, it is crucial that the UK innovation ecosystem encourages sustainable and responsible business practices which aim to avoid**

conflicts of interests (e.g., platforms abusing their control of marketplaces, using dark patterns or profiting from spread of harmful content) and minimise the risks of immersive technologies to create and exacerbate individual and social harms in various domains of life including work, education, health, entertainment, etc.

Responsible Research Strategy

The Responsible Innovation strand highlights the need to link academic research with industry. As a priority, researchers should determine what safeguards are needed to improve users' agency, prevent non-consensual capture of data from non-users, and any other harmful, exploitative scenarios. Some of these harms have already been documented by academics and journalists, extending for example to cyberbullying and harassment of vulnerable groups; spread of hate speech and misinformation; user profiling and data harvesting [37]. However, beyond these, an agenda for responsible and sustainable innovation should anticipate the following social and ethical issues: 1) digital escapism and addiction; 2) discrimination, exclusion, and unequal access; 3) lock-in and lack of interoperability; 4) financial speculation and fraud linked with the exchange of digital assets in the Metaverse.

4. Regulatory frameworks

Industry response

Major technology companies (Google, Microsoft, Meta) have begun to build internal capabilities in response to the numerous ethical concerns over emerging digital technologies. In the case of Meta, they have published a list of responsible innovation principles [38] that point to the company's commitment to address issues of privacy, safety, agency, transparency, and inclusivity. While these are positive efforts, government should explore the need for new regulation and policy frameworks to audit the extent to which voluntary responsible innovation pledges by technology companies can address or fail to address individual and social harms linked by the Metaverse [38].

Interoperability standards

A key issue for inclusive innovation is to avoid technology lock-in through incentivising interoperability between (hardware and software) platforms and the lowering of barriers to access which could be due to the cost of hardware or the lack of alternatives. Currently accessing Meta's Metaverse is only possible with a Meta (or its subsidiaries) user account as well as the use of costly hardware. This not only facilitates monopolisation but creates barriers of access which could exclude disadvantaged groups of the population.

Principles- vs rules-based regulations

Recent research from REPHRAIN Policy and Regulation strand researchers explored the advantages and disadvantages of regulatory frameworks concerned with secure adoption of emerging technologies. There is a tension between prescriptive and outcome-based regulations or standards. Prescriptive standards and regulations outline baseline minimum requirements, an approach which is beneficial for stakeholders without previous security expertise who need support in understanding what good level of security provision looks like. However, this prescriptive approach is critiqued for its tendency towards technocratic measures and inflexibility in the face of fast-paced technology development. On the other hand, outcome-based regulations outline ideal high-level principles or security without specifying how to achieve them. They allow a degree of interpretation to suit a given context as technologies evolve and develop. Outcome-based approaches are critiqued for excess subjectivity and difficulties with benchmarking (I.e., understanding what 'good security' looks like across the sector and comparison between organisations) A pre-print [39] is available on demand from Dr Ola Michalec ola.michalec@bristol.ac.uk

5. Policy recommendations

The Metaverse-enabling immersive technologies

Anticipate and address the emergent risks of immersive technologies within the Online Safety Bill: Mass adoption of social media has engendered a host of societal harms (e.g., the impact of social media on political discourse, targeted harassment, or the spread of harmful content, manipulation). Society needs to anticipate and address XR societal harms around manipulation, information disorder, and governance around where, when, and how our view of reality can be altered or augmented - before these harms are realized [21].

We believe that now is an ideal time to regulate immersive technologies and incorporate it in flagship regulatory initiatives like the Online Safety Bill. Fictional virtual spaces that promote the normalisation of radical ideologies (e.g., misogynistic ideals or fascist propaganda) pose a unique risk. Clear oversight of the creation of new virtual spaces is a necessity. We now have a chance to take the anticipatory approach to governance, learn from current debates on online harms in mature digital technologies (e.g., online marketplaces of social media platforms). Such anticipatory approach would allow regulators and developers to understand users' requirements and design appropriate privacy and security safeguards. For example, immersive environments should be transparent about their purpose (workplace collaboration, fictional gaming worlds or advertising).

Harmonise regulatory frameworks on user generated harmful content across the UK and the EU In line with the proposals of the EU Digital Services Act [40], users must be held legally accountable for illegal content that they generate. This accountability can be achieved through ensuring that virtual identities link to real-world identities. Whilst this may present concerns around privacy, this authentication need not be managed by a platform or commercial interest, but instead by an independent regulatory body such as Ofcom. Further, the use of an independent body will enable greater coordination across platforms/metaverses.

Update GDPR to apply to the immersive technologies and enhance users' privacy: The transference and processing of sensitive personal data should be transparent and highly regulated. XR devices lack appropriate user privacy protections, with limited capacity to restrict the data made available to apps. Bystanders have no capacity to consent to how/if they are being captured, sensed, augmented [8].

Mandate safeguarding standards for Social VR: Currently, social VR lacks effective moderation and reporting, and there are no standardised protections on platforms. Examples of basic protection mechanisms could include disabling the setting which allows other users to come within three feet of each other. This simple fix would make individuals less vulnerable to cyber-assaults, as it would prevent users being able to 'touch' each other in a virtual space.

Web 3.0: Blockchain / Decentralised social networks

Clarify guidelines regarding fraud prevention: Financial authorities and regulators (i.e., the Financial Conduct Authority) have provided guidelines to prevent APP frauds occurrence and improve victims' protection, but these guidelines (like the CRM code [4]) are still vague and open to interpretation. This could be achieved by embedding secure and transparent mechanism (like the Payment with Dispute Resolution Protocol) into online banking systems standards.

Include decentralised social media in algorithmic fairness and market regulation frameworks. Government bodies like the Competitions and Market Authority or Ofcom are currently designing policy frameworks aiming to detect and address monopolising effects and exclusionary behaviours in marketplaces and digital services. Emerging technologies, like decentralised social media should come under scope of these initiatives. Special regulatory attention should be given to issues of market manipulation and speculation with the exchange of virtual goods using blockchain and NFTs.

Joint recommendations

Improve literacy around the capabilities and risks of decentralised web and the Metaverse. Authorities like Ofcom and Department of Education should launch a public/school campaign highlighting online harms in these emerging technologies and ways to prevent or minimise them. The campaign ought to include harms to bystanders/non-users and the differences in functionality of platforms vs protocols.

Support responsible innovation collaborations across academia and industry: Research Councils ought to encourage the creation of academic-industry-policy networks focusing on responsible adoption of emerging digital technologies. These networks would focus on data sharing across organisations, API access, and improving accountability of R&D processes.

Ensure equitable access to emerging technologies and tackle risks of digital exclusion. There is a risk that marginalised groups (e.g., those with physical disabilities such as vision impaired as well as low-income citizens) won't be able to benefit from immersive technologies or Web 3.0 services. Civic and democratic bodies (e.g., local councils) should not employ the Metaverse and Web 3.0 until these concerns have been addressed.

References

[1] Mathis, Florian, Hassan Ismail Fawaz, and Mohamed Khamis (2020) "Knowledge-driven biometric authentication in virtual reality." In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-10.

[2] Mathis, Florian, John H. Williamson, Kami Vaniea, and Mohamed Khamis (2021) "Fast and secure authentication in virtual reality using coordinated 3d manipulation and pointing." ACM Transactions on Computer-Human Interaction (ToCHI) 28, no. 1: 1-44.

[3] UK Finance (2021) UK Finance Report https://www.ukfinance.org.uk/system/files/Half-year-fraud-update-2021-FINAL.pdf

[4] HM Treasury (2022) Government approach to authorised push payment scam reimbursement - The Contingent Reimbursement Model Code https://www.gov.uk/government/publications/government-approach-to-authorised-push-payment-scam-reimbursement/government-approach-to-authorised-push-payment-scam-reimbursement

[5] Abadi, Aydin, Steven J. Murdoch, and Thomas Zacharias (2022) "Recurring Contingent Service Payment." *arXiv preprint arXiv:2208.00283*

[6] Abadi, Aydin, and Steven J. Murdoch (2022) "Payment with Dispute Resolution: A Protocol For Reimbursing Frauds' Victims." *Cryptology ePrint Archive*

[7] Campanelli, Matteo, Rosario Gennaro, Steven Goldfeder, and Luca Nizzardo (2017) "Zero-knowledge contingent payments revisited: Attacks and payments for services." In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 229-243.

[8] REPHRAIN National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (2023) The Map of Online Harms, Risks, and Vulnerabilities. https://rephrain-map.co.uk/

[9] Cork, A., Smith, L. G. E., Ellis, D. A., Stanton Fraser, D., & Joinson, A. (2022) Rethinking Online Harm: A Psychological Model of Contextual Vulnerability. Pre-print. https://doi.org/10.31234/osf.io/z7re2[38]

[10] Barreda-Ángeles, Miguel, and Tilo Hartmann (2022) "Psychological benefits of using social virtual reality platforms during the covid-19 pandemic: The role of social and spatial presence." *Computers in Human Behavior* 127: 107047.

[11] Riva, Giuseppe, Luca Bernardelli, Matthew HEM Browning, Gianluca Castelnuovo, Silvia Cavedoni, Alice Chirico, Pietro Cipresso (2020) "COVID feel good—an easy self-help virtual reality protocol to overcome the psychological burden of coronavirus." *Frontiers in psychiatry* 11: 563319.

[12] Matei, Sorin, and Sandra J. Ball-Rokeach (2001) "Real and virtual social ties: Connections in the everyday lives of seven ethnic neighborhoods." *American Behavioral Scientist* 45, no. 3: 550-564.

[13] VRC Chat (2023) Steam charts data https://steamdb.info/app/438100/graphs/

[14] Maloney, Divine, Guo Freeman, and Andrew Robb (2021) "Stay connected in an immersive world: Why teenagers engage in social virtual reality." In *Interaction Design and Children*, pp. 69-79.

[15] Maloney, Divine, Guo Freeman, and Andrew Robb (2020) "A virtual space for all: Exploring children's experience in social Virtual Reality." In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pp. 472-483.

[16] Freeman, Guo, Samaneh Zamanifard, Divine Maloney, and Dane Acena (2022) "Disturbing the peace: Experiencing and mitigating emerging harassment in social virtual reality." *Proceedings of the ACM on Human-Computer Interaction* 6, no. CSCW1: 1-30.

[17] Blackwell, Lindsay, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz (2019) "Harassment in social virtual reality: Challenges for platform governance." *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW: 1-25.

[18] Fiani, Cristina, Pejman Saeghe, Mark McGill, Mohamed Khamis (2023). Parent and Adult Perspectives on Children's Use of Social Virtual Reality. Currently in Review for ACM CSCW 2023 https://cscw.acm.org/2023/

[19] Fiani, Cristina, Robin Bretin, Mark McGill, Mohamed Khamis (2023). Big Buddy: Exploring Child Reactions and Parental Perceptions towards a Simulated Embodied Moderating System for Social Virtual Reality. Currently in Review for ACM IDC 2023 https://idc.acm.org/2023/

[20] McGill, Mark (2021) Extended Reality (XR) and the Erosion of Anonymity and Privacy. *The IEEE Global Initiative on Ethics of Extended Reality (XR) Report*. (Nov. 2021), 1–24.

[21] Abraham, Melvin, Pejman Saeghe, Mark Mcgill, and Mohamed Khamis (2022) "Implications of XR on Privacy, Security and Behaviour: Insights from Experts." In *Nordic Human-Computer Interaction Conference*, pp. 1-12.

[22] Gugenheimer, Jan, Wen-Jie Tseng, Abraham Hani Mhaidli, Jan Ole Rixen, Mark McGill, Michael Nebeling, Mohamed Khamis, Florian Schaub, and Sanchari Das (2022) "Novel Challenges of Safety, Security and Privacy in Extended Reality." In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1-5.

[23] Tseng, Wen-Jie, Elise Bonnail, Mark Mcgill, Mohamed Khamis, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. (2022) "The dark side of perceptual manipulations in virtual reality." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1-15.

[24] O'Hagan, Joseph, Julie R. Williamson, Mark McGill, and Mohamed Khamis (2021) "Safety, power imbalances, ethics and proxy sex: Surveying in-the-wild interactions between vr users and bystanders." In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 211-220. IEEE

[25] Bonnail, Elise, Wen-Jie Tseng, Mark McGill, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer (2023) "Exploring Memory Manipulation in Extended Reality Using Scenario Construction." *To Appear In CHI Conference on Human Factors in Computing Systems*

[26] Saeghe, Pejman, Mohamed Khamis, and Mark McGill. (2023) PriXR: Information Disorder, User Manipulation, Safety, and Privacy: Addressing the Vulnerabilities and Harms of Everyday Extended Reality. *REPHRAIN briefing*

[27] Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. European Review of Social Psychology, 32(2), 348-384.

[28] O'Hagan, Joseph, Pejman Saeghe, Jan Gugenheimer, Daniel Medeiros, Karola Marky, Mohamed Khamis, and Mark McGill (2023) "Privacy-Enhancing Technology and Everyday Augmented Reality: Understanding Bystanders' Varying Needs for Awareness and Consent." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, no. 4: 1-35.

[29] Lemley, Mark A., and Eugene Volokh (2018) "Law, virtual reality, and augmented reality." *University of Pennsylvania Law Review*: 1051-1138.

[30] Kent Bye (2019) XR Ethics Manifesto.

[31] Krauß, V., Saeghe, P., Boden, A., Khamis, M., McGill, M., Gugenheimer, J. and Nebeling, M (2023) What Makes XR Dark? Examining Emerging Dark Patterns in Augmented and Virtual Reality through Expert Co-Design. *In Review for ACM ToCHI*

[32] Hassan, Anaobi Ishaku, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson (2021) "Exploring content moderation in the decentralised web: The pleroma case." In Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies, pp. 328-335.

[33] Ishaku Anaobi, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Damilola Ibosiola and Gareth Tyson (2023) "With Great Power comes Great Responsibility: Exploring Administration in Decentralized Social Networks." In Proceedings of the Web Conference.

[34] Bin Zia, Haris, Aravindh Raman, Ignacio Castro, Ishaku Hassan Anaobi, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson (2022) "Toxicity in the decentralized web and the potential for model sharing." Proceedings of the ACM on Measurement and Analysis of Computing Systems 6, no. 2: 1-25.

[35] Hindman, Matthew (2018) *The Internet Trap: How the Digital Economy Builds Monopolies and Undermines Democracy*. Princeton University Press.

[36] The Guardian (2021) Metaverse is just a new venue for the age-old problem of sexual harassment https://www.theguardian.com/commentisfree/2021/dec/18/metaverse-new-venue-sexual-harassment-facebook

[37] Meta (2022) Responsible Innovation starts with privacy. https://about.facebook.com/metaverse/responsible-innovation/

[38] Elettra Bietti (2021) “From Ethics Washing to Ethics Bashing: A Moral Philosophy View on Tech Ethics,” *Journal of Social Computing* 2, no. 3; 266–83, https://doi.org/10.23919/JSC.2021.0031

[39] Michalec, Ola, Ben Shreeve and Awais Rashid (2022). Cyber Security Visions of Future Energy Systems: Design, Support Function or Public Trust? PETRAS Academic Community Conference 16-17th June 2022. Available at: https://petras-iot.org/project/understanding-disruptive-powers-of-iot-in-the-energy-sector-power2/

[40] European Commission (2022) The Digital Services Act: ensuring a safe and accountable online environment. Strategy. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en