BEYOND THE HEADSET

Charting a course for safer experiences for children in extended reality environments

ROUNDTABLE SUMMARY PAPER
August 2024

Contents

Foreword **2**
Key risks and issues **3**
Easy access to children and user anonymity 3
New methods to create and stream abuse 3
Creating content and accessing XR 4
Links to other forms of victimisation 4
Current solutions and gaps in the response **5**
Content moderation 5
Age assurance and user authentication 5
Safety by design 6
Bystander intervention 6
Emerging themes for further exploration **7**
Engaging with children and young people to better understand their needs 7
Intersection with artificial intelligence (AI) 7
Policies and regulations 8
Conclusion **9**

Roundtable attendees

- African Society for Cyber Security Awareness
- Ministry of Justice and Public Security, Brazil
- Child Focus
- ChildFund International
- ECPAT International
- eSafety Commissioner, Australia
- Faculty AI
- Film and Publication Board of South Africa
- Home Office, UK Government
- Korea Communications Standards Commission
- Meta
- National Police Chiefs' Council (UK)
- Resolver
- SafeToNet Foundation
- Tech Coalition
- University of Middlesex
- WeProtect Global Alliance

Key terminology

Extended Reality (XR) serves as a comprehensive term for immersive technologies that merge the physical and virtual worlds. This category encompasses Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR).

Virtual Reality (VR) fully immerses users in a completely virtual environment. Users typically wear headsets that display a simulated world, effectively isolating them from the real world. Interaction within this virtual space is facilitated through controllers or hand-tracking systems.

Augmented Reality (AR) enhances the real-world environment by overlaying digital elements. This technology "augments" the user's view with additional information or virtual objects, thereby enriching their perception of the real world.

Mixed Reality (MR) integrates elements of both VR and AR, enabling digital objects to interact with the real-world environment. This creates a blended experience where virtual and physical elements coexist and interact seamlessly.

Foreword

Extended reality (XR) technologies have existed for decades, but recent years have seen a surge in interest and availability in workplaces and consumer leisure sectors. Established technology companies have invested heavily in XR, with the global market forecasted to surpass $1.1 trillion by 2030. As XR becomes more accessible, offenders are increasingly likely to exploit these technologies to sexually abuse and harm children.

WeProtect Global Alliance's 2023 Intelligence briefing on XR technologies and child sexual exploitation and abuse highlighted the risks XR poses to child safety, including opportunities for offenders to access victims, distribute child sexual abuse material, simulate abuse of virtual children, and use haptics to mimic real-world sensations like movements and force. The privacy of VR environments, where users are isolated by headsets, further exacerbates these risks.

Interoperability blurs accountability for user safety as users move seamlessly between immersive environments.

Although there is currently limited evidence of the use of XR in child sexual exploitation and abuse, identified risks are wide-ranging and include:

- opportunities for offenders to access children
- distribution of child sexual abuse material
- simulated abuse of virtual representations of children
- abuse of children facilitated by integrated technologies such as haptics.

Children make up a large portion of the XR user base. According to research from the Bracket Foundation and Value for Good, from 2016 to 2021 the gaming industry's global revenues grew by more than 50% to reach $198.4 billion in 2021/22 – roughly four times the size of the global digital home entertainment market.

Gaming has established itself as part of many children's daily lives. According to the Entertainment Software Association 2024 report, 61% of the United States population reports playing video games for over an hour a week with 5-18 year olds representing 24% of the user base. Children in the United States are reported to game on average for 2.5-3 hours a week, with those under eight spending an average of 23 minutes gaming daily.

However, geographical, social, economic and gender disparities exist around access to XR. Extended reality technology is out of the reach of many children around the world because of high-cost thresholds, limited access to high-speed internet and the computer and technological hardware required to run this technology.

We are currently at a tipping point which necessitates that policymakers and industry ensure the safe use of XR technologies by children.

In May 2024, WeProtect Global Alliance held a roundtable with members from technology companies, child safety organisations, policy makers, educators and researchers from across the world to explore XR threats and how we might improve responses. Their combined expertise provided a comprehensive view of the opportunities provided by this technology, and the challenges in protecting children from risk of harm.

This report summarises key themes and insights from this discussion, focusing on potential harms in XR environments and current mitigations, including robust content moderation, age assurance, user authentication, Safety by Design and collective social responsibility.

It includes actionable strategies to create a safer XR environment for children. It emphasises the need for continued collaboration among all stakeholders to ensure XR technologies are safe and enriching for young users.

The Alliance would like to thank Professor Elena Martellozzo (Professor in Criminology and Associate Director at the Centre for Child Abuse and Trauma Studies at Middlesex University), David Miles (Director of Safety Policy for Europe, Middle East and Africa at Meta) and Ian Critchley QPM (UK National Police Chiefs Council Lead for Child Protection, Abuse and Investigation and member of the National Child Safeguarding Practice Review Panel) for their respective presentations on technology, children's rights and law enforcement. We would also like to thank all participants in the roundtable discussion for their candour, insights and expertise.

Key risks and issues

Participants discussed key risks and issues ranging from methods used to stream abuse, links to other forms of victimisation and pathways to further offending.

There was a lack of consensus and common understanding or definition of XR. Some associated it solely with virtual reality (VR) headsets while others felt XR encompasses a broader spectrum. The conversation also reflected differences in opinion and understanding of the technology. Some participants felt it affected a limited population with very few current users, while others felt it encompassed a broader range of digital technologies including video calls such as Zoom or Microsoft Teams. This ambiguity complicates the ability to develop a coordinated and consistent response.

Easy access to children and user anonymity

Many XR platforms allow for relatively anonymous user profiles, enabling offenders to pose as fellow gamers or community members while concealing their real identity. This anonymity, combined with the open access of these platforms, makes it easier for offenders to contact and exploit children.

Abusers might use XR to initiate child grooming or entice victims into progressively more harmful forms of exploitation. Once initial trust or rapport is established, offenders can introduce victims to other platforms where further abuse can occur.

Participants reflected that XR environments can be difficult to monitor due to the multitude of ways in which users can interact with each other. XR provides predators with an avenue to build trust and connections with children with promises of friendship, special access or exciting content, before moving the communication away from XR to messaging apps or social media. This allows predators to establish contact outside XR safety measures and indicates a need for a robust response across all digital environments.

> ***We know that where technology goes, abuse follows. Offenders will continue to adapt their behaviours to identify where children are. Look at the stats – we've seen a 400% increase in reporting. That's not a good place to be, and we can prevent a lot of it. –*** *participant*

It was noted that different types of online sexual violence are gendered, with boys at particular risk in online gaming environments. They are also likely to face stigma and social barriers which may contribute to underreporting of sexual exploitation. It was considered important to better understand and leverage the technologies popular with young boys to address risks and devise support mechanisms. A helpful catalyst to encouraging behavioural change may also be to think about positive framing around how to be safe, and to explore ways technology could be leveraged to support prevention.

New methods to create and stream abuse

The immersive nature of XR technologies allows abusers to create and stream realistic environments where children can be exploited. Virtual environments can simulate real-world locations or social scenarios, often making victims feel as if they're engaging in a safe, trustworthy space.

Participants highlighted that behavioural patterns of offenders in XR environments are also deployed in gaming and social media environments. Predators can exploit 'off-platforming', the act of moving communication from XR environments to other platforms, to groom children. Suojellan Lapsia's recent study 'Tech platforms used by online child sexual abuse offenders' was cited as evidence of perpetrators establishing contact with children on social media or gaming platforms and then moving to more private environments, often end-to-end encrypted messaging applications, where evidence of grooming is harder to detect, report or disrupt.

Extended reality technologies featuring hardware such as headsets can also be isolating to users. VR headsets fully immerse users in a virtual world, often isolating them from real-world supervision or intervention. This lack of external awareness can prevent children from recognising grooming behaviours or seeking help in time.

Presentations to the group also flagged concerns about blurred lines between virtual and physical spaces. Advanced technologies like haptics simulate real-world sensations, further enhancing immersion and realism. Haptic technology focuses on vibrations, textures, and light pushes to simulate realistic experiences. The potential negative physical and psychological effects if these are misused to exploit children can result in very real harm to children.

Finally, perpetrators evade detection by using blockchain technology to cover their tracks. Blockchain can be used to anonymise and encrypt transactions, making it harder to trace the identities and activities of offenders who use XR to distribute child sexual abuse material.

Links to other forms of victimisation

Offenders can use XR environments to manipulate or groom children through tailored interactions. By gaining their trust through immersive social spaces, predators can lure victims into exploitative behaviours or other harmful forms of victimisation. Research by Child Focus indicates that XR increases risks of victimisation, live deepfakes, online sexual violence, and harassment for children.

XR technology was described as an environment enabling predators to 'throw out more lines to catch more fish'. In-app currencies, particularly in gaming environments, were also highlighted as a potential incentive or triggering factor in persuading children to engage in dangerous online behaviour.

Participants noted different tactics being used to manipulate children from offenders 'live deep faking' themselves in order to sound and look like young girls, to using extortion and sharing grooming scripts in a diverse range of languages.

Sexual extortion or 'sextortion' was highlighted as having 'gone through the roof' in the past year and was underlined as a key concern as XR technologies continue to expand their market footprint. The National Centre for Missing and Exploited Children (NCMEC) – which acts as the global sorting house for online sexual harms against children – recently reported that cases of sexual extortion have doubled in a year.

> ***We know perpetrators will find any crack in platforms they can penetrate. There's a huge amount of work to be done, but also a lot of knowledge we already have. –*** *participant*

Creating content and accessing XR

Most consumer content is accessed via app stores controlled by platform providers. The content available on any given platform depends on what the app store allows. The lack of interoperability between platforms may hamper developers and reduce availability as the development effort required to make different versions of an app for different platforms may not make economic sense.

Consumer AR content for mobile phones is widely available in app stores like Google Play and the Apple App Store. Industrial and workplace AR applications including those for more sophisticated AR viewers like Microsoft HoloLens may be bespoke developments for specific customers or sectors.

These stores have content rules designed to keep out illegal and harmful content. Android phones do not require the use of the Google Play store, with other stores being available such as from Amazon and the opensource F-Droid repository, along with supporting manually installed applications.

VR content is available via several different platforms. Some software platforms are entirely controlled by a single company, such as Sony's PlayStation VR platform. As with AR content, applications must meet requirements for acceptable content and pass a review before publication.

Current solutions and gaps in the response

Participants discussed current responses to XR risks and suggested further solutions. A strong prevention approach and adoption of Safety by Design[1] was supported with the opportunity to make XR safer before there is wider global adoption of this technology.

Content moderation

As with social media, participants agreed community standards should apply in XR environments, and be enforced by content moderation teams.

Content classifiers were highlighted as a useful technical tool in detecting child sexual exploitation and abuse in XR environments. These machine learning algorithms are trained on vast amounts of data to identify patterns indicative of child sexual abuse in XR environments. Classifiers are beginning to be able to analyse images, videos, and even signals from user behaviour to flag potential abuse. The types of content that these systems can detect include and are not exclusive to recognising nudity, sexualised poses, and sexually suggestive content.

The multitude of XR enabled capabilities, including audio, speech, live video and other forms of content creation, make it harder to moderate content in the same ways as traditional social media (e.g. hash matching). This is further exacerbated by the fact that interactions are live and take place in real time. Some organisations reported applying the same approaches of sophisticated AI classifiers to XR apps and environments that they have been developing to tackle child sexual exploitation and abuse on live streaming platforms.

It was noted that classifiers still have limitations – from "false positives" where content is misidentified to struggling to identify new or uncommon types of abuse. Participants supported a multi-layered approach to content moderation.

> ***We should be thinking about content moderation and child protection... getting on the front foot with XR environments and ensuring we're thinking about it from the start. –*** *participant*

It was also felt that content flagging systems deployed in XR need to be workable across platforms and service providers, due to the diversity of services provided in XR and the fact that offenders often move across platforms to hide their tracks. The Tech Coalition's Lantern project (the first cross-platform signal sharing programme) was highlighted as a good example of a collaborative approach to tackling this. Participants felt a way to identify and alert platforms when known bad actors shift from one service to another would be useful.

Large language models (LLMs) were mentioned as a key technology being used to enforce community standards in XR spaces. The hashing of voice chats is also being deployed to use as a kind of digital receipt for a voice chat that has taken place and could help with evidence collection.

While these AI tools are important for detecting illegal behaviour in extended reality environments, participants felt they should be deployed in combination with other tools and systems such as trusted flaggers, user reporting, live operator interventions and well-resourced content moderation teams.

Age assurance and user authentication

Perpetrator and victim identification are significant challenges for law enforcement authorities as they try to protect children and identify offenders.

Age assurance and user authentication were underlined as vital safeguards for protecting children in XR and metaverse environments. Verifying a user's age helps children to explore age-appropriate online environments, while preventing them from encountering inappropriate content or interacting with malicious users.

At the same time there is an opportunity to reduce ease of access to children for grooming, exploitation and abuse by perpetrators and those at risk of offending by ensuring that children have their own, safe places to learn, play and interact with known friends, peers and people their own age.

Participants felt anonymity facilitated unwanted interactions and potential grooming. In much the same way as people need to provide ID when buying alcohol or entering a nightclub, participants felt we should be able to ask the same of users in the virtual world. User authentication ensures that only authorised individuals can access the XR or metaverse space, limiting the potential for harm and facilitating the gathering of digital forensics by law enforcement.

The combination of these measures creates a safer and more controlled environment for children to explore the wonders of XR and the metaverse.

1 – A 'Safety by Design' approach includes assessing the impact of all products and services from a child rights perspective. Online service providers should identify and, as appropriate, warn, expel and report actors who pose a risk to children (Global Threat Assessment 2021 and eSafety Safety by Design).

Safety by design

Participants reflected there is already a demonstrable amount of knowledge about the nature of online sexual offending, offender behaviours and how they operate. However, this knowledge needs to be better used to inform safety by design, security by design and safeguarding by design of emerging technologies. Safety by design requires technology companies to consider how to minimise threats and harms throughout the design, development and deployment process.

In addition to age assurance and user authentication techniques, participants cited features like default parental controls, age-appropriate content filters, personal boundary settings, teenage user controls and time limits as preventative tools. There is a wealth of young start-up companies in the safety tech sphere working on developing and providing these varied solutions.

Relying solely on parental controls alone was considered an inadequate response because parents are unable to adapt with the same speed and agility to new technologies as their children. By integrating a combination of these safeguards from the very beginning, XR and metaverse experiences can be designed to nurture children's creativity and development while minimising potential risks like addiction, cyberbullying and exposure to inappropriate content.

While many current safety by design measures are used to prevent and disrupt sexual exploitation and abuse perpetrated by adults against children, participants also noted the increased prevalence of 'self-generated' images and peer-to-peer abuse, which need to be considered further in safety by design measures.

It was suggested that a clear and prescriptive approach should be developed on how extended reality technologies consider safety by design.

> ***People, parents, carers, law enforcement outside the regulatory regime all have a responsibility. There's also a moral responsibility from tech companies to speak about it publicly.*** – *participant*

> ***There's no separation between the online environment and the community we live in. Violence against women, bullying – they are all increasing. Behaviour trumps everything - the way we interact with people [on and offline] matters.*** – *participant*

Bystander intervention

There was strong support for collective and social responsibility for ensuring children's safety in XR environments.

Bystander intervention, where platform users in XR spaces are both vigilant and actively encouraged to report concerning behaviour or content, acts as a vital second line of defence. Ensuring that users can intervene and support other users when they see wrongdoing is crucial to building a culture of respect in these new digital spaces. Reporting tools need to be easily located, accessible and user-friendly to ensure that users use the reporting tools. This in turn empowers users to flag potential predators, abusive interactions, or inappropriate content exposure.

Platforms can encourage bystander intervention through clear reporting mechanisms, educational campaigns, and fostering a sense of community responsibility. By working together, users can create a safer and more positive XR experience for everyone, especially for children who may be hesitant or unaware of how to report issues themselves.

Participants felt that while this guidance would be useful to all users, the responsibility of ensuring a safe online space should not fall on users alone, and they should not be depended upon to police online spaces. Bystander intervention should be encouraged alongside a fully implemented safety by design approach.

Emerging themes for further exploration

Engaging with children and young people to better understand their needs

WeProtect Global Alliance's 2023 Global Threat Assessment highlighted the urgent need for child-centred approaches, with participation and consultation essential tools to empowering children and young people to contribute to discussions and decisions that affect them.

Upcoming research presented at the roundtable reflected children are excited to explore 3D worlds and feel the metaverse is more real than 2D online games. Reality confusion and encountering people they don't know were highlighted as the biggest risks as children felt safer in an XR environment where other people they know are present.

The research found that when children were asked what they need to feel safe in the metaverse, they cited increased content moderation and behaviour guidance as the top key priorities, followed by time restrictions and being around adults during use. Education about risks and harm in these new online environments was highlighted as the first line of defence. Talking to children about the risks online and teaching them to 'identify the red flags' needs to remain a priority, along with positive prevention messaging.

A potential solution suggested was to provide children with a sliding scale of accessibility where they should be able to use XR for activities that are age-appropriate or beneficial to them (e.g. at school) but with limited access to more sensitive content depending on user age. On this point, it was emphasised that having clear regulation and guidance for implementing this would be helpful.

Participants also reflected on geographical and societal disparities contributing to a gap in XR accessibility and recognising that not all children's experiences will be the same. The high cost of XR headsets and supporting devices puts this technology out of reach for many children, particularly those in the Global South or low-income households. This lack of familiarity with the technology could hinder children's ability to identify threats and develop strong digital media literacy competencies within XR environments.

Moving forward we need to ensure equitable access to XR technology for children around the world, while also implementing targeted and localised digital literacy programmes for XR that are relevant to the issues and challenges that children in different countries or regions face.

> ***I just can't see any way we can have meaningful success with AI without improving user authentication and age verification. If we see XR as an extension of the real world, and we can do [identity authentication] in the real world, we can also do it in the virtual world.*** – *participant*

Intersection with artificial intelligence (AI)

Discussion focussed on how the intersection of AI and the metaverse creates a unique set of risks for online child safety, such as those outlined by the Australian eSafety Commissioner in their Generative AI position statement. For example, AI can be used to create highly realistic and disturbing content, including deepfakes, that could be used to produce CSAM featuring children or avatars that resemble children. This content can also potentially be produced on an industrial scale.

The discussion noted that despite the recent acceleration of interest in this space, we are still at the formative phase of generative AI and AI-generated child sexual abuse material. XR technologies are also still under development, so the full scope of the risks and the intersection with AI is not yet fully understood.

What was clear from the discussion is that many harms which depend on both technologies are emerging including live deep fakes, AI-powered grooming chats (via chatbots) and non-consensual virtual contact. AI-generated content is also becoming more photorealistic.

Participants felt more needed to be done to help content moderators, law enforcement authorities and members of the public to identify AI generated content.

They also felt that content moderation at the specific intersection of AI and XR requires a deep level of knowledge, understanding and training in a sector where employee turnover is high. Strong onboarding and robust training must be provided to content moderators – as well as frontline workers in policing and hotlines – so they can identify behaviours, synthetic content, know what content to look for and understand how to moderate content in XR environments.

Overall, the roundtable reflected that tackling the harms at the intersection of XR and AI requires a multi-pronged approach encompassing technological advancements, legal frameworks, user awareness, and industry responsibility.

Policies and regulations

It was acknowledged during the discussion that lots of progress has been made regarding online safety regulation and there was a sense of optimism in the strides made globally to establish aligned rules, regulations and frameworks. Collaboration between regulators is fostering a movement towards alignment – for example, via the Global Online Safety Regulators' Network (GOSRN) – even with the challenges of balancing diverse national approaches and systems.

Participants shared how they were working to ensure that existing rules and laws can be implemented to address the challenges posed by XR by placing obligations on platforms to protect children and taking action to report illegal and problematic content to hotlines or law enforcement authorities.

They also reflected on the need for governments and regulators to continually revise and update pre-existing laws and policies to keep pace with technologies and new emerging threats. It was felt that the private sector, governments and regulators needed to work together to establish clear rules and best practices in areas such as XR content moderation, age verification, parental controls and other safety risk mitigation strategies.

> ***On the regulatory side I see more progress being made and also more collaboration across the world. –*** *participant*

The roundtable reflected that governments and regulators are constantly navigating and learning how to embrace new technology. For example, regulators are seeing avatars being used to exert threatening behaviour in extended reality environments but are finding it very difficult to address incidents in XR or the metaverse when an avatar is harassed or abused as this is rarely covered under existing laws or regulations.

Participants discussed the need to provide law enforcement with the tools and resources to adapt and develop new strategies to effectively protect children in XR. This requires a multi-pronged approach from standardised guidance for detecting and reporting illegal content and behaviour to specialised training programs for law enforcement officers to equip them with the skills to investigate crimes in virtual environments. Digital forensics and evidence gathering skills were considered critical to investigating these new realities.

Conclusion

The roundtable brought together global stakeholders from different sectors to develop a common understanding of how we can tackle child sexual exploitation and abuse in extended reality environments that transcend borders.

On the positive side, children are excited about this emerging technology and how they can learn, explore and play in these new digital spaces. However, technology also provides new avenues for predators to sexually abuse and exploit children. Issues such as grooming, verbal and physical abuse, sexual extortion, live-deepfakes and AI-generated child sexual abuse material are all high-priority concerns for XR experiences.

There was strong support across sectors to tackle the more challenging aspects of this technology in a globally aligned way. By prioritising proactive risk assessments and increasing preventative measures, we can improve the mitigation of these emerging risks.

Areas for focus include:

- Ensuring content moderation systems deployed in XR are workable across platforms due to the diversity of services provided in XR and the fact that offenders often move across platforms to hide their tracks.
- Develop a comprehensive approach to identify illegal activity in XR, blending tools such as AI with human oversight through trusted flaggers, user reporting, live interventions and robust content moderation teams.
- Ensuring user safety, user authentication systems and age assurance mechanisms are implemented to address the challenges raised by user anonymity, reduce potential harm and allow law enforcement to gather digital evidence more easily.
- Support equitable access to technology for all children to bridge the digital divide in XR.
- Provide targeted digital literacy programmes, developed in collaboration with children to address the real challenges they face.
- Equip content moderators and frontline workers in law enforcement and hotlines with comprehensive training to identify harmful behaviours, recognise suspicious content (including AI-generated material), and effectively moderate XR environments.

This is a tipping point. We are at a unique moment in time where there is potential to mitigate the risk of harm to children. While harms against children are not currently being reported on the same scale as other online harms across social media, gaming and live-streaming environments, as XR technology become more available, we have a window of opportunity to shape current and new technologies to ensure that they are safe for children from the start.

Further reading and resources

- 2023 Global Threat Assessment – WeProtect Global Alliance
- Extended reality technologies and child sexual exploitation and abuse – University of Manchester & WeProtect Global Alliance
- Model National Response Framework – WeProtect Global Alliance
- Safeguarding Children in the Metaverse – REPHRAIN National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online
- The Metaverse, Online Sexual Exploitation and Sexual Abuse of Children – a new challenge for today's global society? – Katarzyna Staciwa, NASK
- The Future of eXtended Reality Technologies, and Implications for Online Child Sexual Exploitation and Abuse – Professor Stephen Pettifer, Professor Emma Barrett, Dr James Marsh, Ms Kathryn Hill, Dr Polly Turner, Dr Sandra Flynn – University of Manchester
- Safety by Design for Generative AI: Preventing Child Sexual Abuse – Thorn and All Tech is Human
- Gaming and the Metaverse – Bracket Foundation and Value for Good in cooperation with the Centre for Artificial Intelligence and Robotics at the United Nations Interregional Crime and Justice Research Institute (UNICRI)
- Children's Views on Gaming – Boston Children's Digital Wellness Lab
- Age assurance trends and challenges – issues paper – eSafety Commissioner, Australian Government
- Generative AI – position statement – eSafety Commissioner, Australian Government
- Immersive technologies – position statement – eSafety Commissioner, Australian Government